# Can Large Language Models Infer Causal Relationships from Real-World Text?

Ryan Saklad[1], Aman Chadha[*2], Oleg Pavlov[1], and Raha Moraffah[1]

[1]Worcester Polytechnic Institute
[2]Amazon Gen AI

## Abstract

Understanding and inferring causal relationships from texts is a core aspect of human cognition and is essential for advancing large language models (LLMs) towards artificial general intelligence. Existing work primarily focuses on synthetically generated texts which involve simple causal relationships explicitly mentioned in the text. This fails to reflect the complexities of real-world tasks. In this paper, we investigate whether LLMs are capable of inferring causal relationships from real-world texts. We develop a benchmark drawn from real-world academic literature which includes diverse texts with respect to length, complexity of relationships (different levels of explicitness, number of events, and causal relationships), and domains and sub-domains. To the best of our knowledge, our benchmark is the first-ever real-world dataset for this task. Our experiments on state-of-the-art LLMs evaluated on our proposed benchmark demonstrate significant challenges, with the best-performing model achieving an average F1 score of only 0.477. Analysis reveals common pitfalls: difficulty with implicitly stated information, in distinguishing relevant causal factors from surrounding contextual details, and with connecting causally relevant information spread across lengthy textual passages. By systematically characterizing these deficiencies, our benchmark offers targeted insights for further research into advancing LLM causal reasoning.

## 1 Introduction

The ability to identify and understand causal relationships embedded within texts is a fundamental aspect of human cognition (Pearl, 2009; Harari, 2014) and is crucial for complex decision-making. Humans excel at inferring these relationships when given a text, even when they are not explicitly stated; for instance, from "Raising prices might

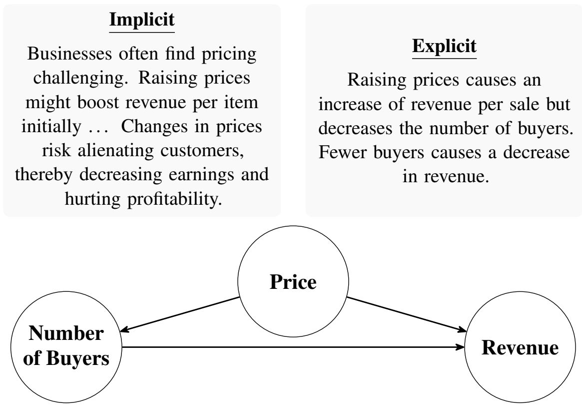


Figure 1: An example causal graph illustrating fundamental economic relationships, showcasing how *Price* affects *Number of Buyers*, which in turn influences *Revenue*. The text example shows two example texts for the same graph, which highlights a challenge in constructing causal relationship purely from text and underscores the need to detect both explicit and implicit cues for accurate causal graph generation.

boost revenue per item initially... Changes in prices risk alienating customers" humans can infer that raising the "*Price*" causally influences the "*number of customers*". Large language models (LLMs) have shown remarkable progress, fueling aspirations towards artificial general intelligence (AGI). Given the importance of causality to intelligence, significant research has focused on assessing the causal reasoning capabilities of LLMs. Much of this work evaluates LLMs' stored causal knowledge about events (Zhou et al., 2024) rather than their ability to infer relationships from textual contexts. While some studies examine extracting causal relationships from text (Veldhuis et al., 2024; Hosseinichimeh et al., 2024; Oh, 2025; Jin et al., 2024; Joshi et al., 2024a; Lasheras and Pinheiro, 2025), they often use synthetically generated or simplified texts with explicitly stated causal links. This approach, however, falls short of real-world scenarios where causal relationships are often embedded in long, complex texts, with varying de-

[*]Work done outside position at Amazon Gen AI.

grees of explicitness, interconnected causal relationships, and diverse domains. Figure 1 illustrates a toy example demonstrating this contrast: it shows that while a synthetic text may present a clear, explicit statement like "Raising prices causes a decrease in revenue," real-world texts often convey this far more implicitly, requiring moving beyond mere comprehension to extract the causal structure.

To address this gap, we introduce **ReCAST** (Realistic Causal Structure from Text), a novel benchmark drawn from real-world academic literature. ReCAST features diverse texts varying in length, complexity of relationships (including different levels of explicitness, number of events, and causal relationships), and spanning multiple domains and sub-domains, mirroring the challenges of real-world causal reasoning from text. Our experiments on state-of-the-art LLMs using ReCAST demonstrate significant struggles, with the best models achieving an average F1 score of only 0.477. Our results reveal common pitfalls: difficulty with implicitly stated information, distinguishing relevant causal factors from contextual details, and connecting causally relevant information spread across lengthy texts. By systematically characterizing these deficiencies, our benchmark offers targeted insights for further research into advancing LLM causal reasoning.

Our main contributions are as follows:

- We propose, to the best of our knowledge, the first benchmark to evaluate the real-world causal reasoning abilities of LLMs from text.
- ReCAST offers benchmarking with varying text and graph sizes from real-world literature, with extensive filtering and post-processing to ensure quality. We further create a LLM-as-a-Judge for automated, fine-grained evaluation.
- Shows the limitations of LLMs at realistic causal reasoning from text across varied samples, including domains, text length, graph size, and degree of confounding.

## 2 Related Work

**Causal Reasoning with LLMs.** As LLMs have grown in capabilities, significant research has explored their causal reasoning abilities. Existing benchmarks often assess these capabilities through various lenses. Some focus on identifying causal structures from pre-defined variable lists or datasets (as surveyed in (Kıcıman et al., 2024) or parts of CausalBench (Zhou et al., 2024)), largely bypassing complex textual inference. Other benchmarks utilize textual inputs, but frequently rely on synthetic or simplified narratives with explicit, often pairwise, relationships or for constructing relatively small graphs (e.g., CLadder (Jin et al., 2024), ExpliCa (Miliani et al., 2025), From text to map (Hosseinichimeh et al., 2024)). Other prior work evaluates causal understanding via question-answering, probing LLMs' stored knowledge and their ability to deduce specific causal claims from short prompts or contexts (e.g., CaLQuest.PT (Lasheras and Pinheiro, 2025), COLD (Joshi et al., 2024a)). While some studies engage with real-world texts, they target more constrained tasks such as identifying a single primary causal link (Oh, 2025) or classifying individual sentences for causality (Veldhuis et al., 2024), rather than requiring the synthesis of a complete causal graph. These approaches, detailed further in Appendix A, generally do not address the challenge of causal graph construction under the full diversity and complexities of real-world conditions, where causality is often implicit and embedded within long narratives. ReCAST differs from prior work by directly targeting this gap, using academic literature as verified source data for causal reasoning from text under realistic conditions.

**Knowledge Graph Discovery with LLMs** Recent years have witnessed substantial progress in knowledge graph construction frameworks with LLM-based approaches (Bratanic, 2024; Zhang and Soh, 2024; Yu et al., 2023). These methods convert unstructured text into structured formats by identifying entities, relationships, and additional attributes. However, they emphasize open-domain and factual information, differing from the unique challenges of causal reasoning.

## 3 The ReCAST Dataset

In order to ensure that ReCAST is a realistic causal reasoning benchmark, we need benchmark samples that have a source text, a ground-truth causal graph, and meet our realism criteria. The construction of the benchmark is achieved via a 3-step pipeline: (i) Paper collection and filtering, (ii) Annotation, (iii) Post-processing. Next, we elaborate on each step.

### 3.1 Paper Collection and Filtering

We focus ReCAST on the economics domain due to the unique characteristics of some of its causal literature, which is especially well-suited for this benchmark. These papers are ideal for a text-based causal benchmark, as they contain detailed textual descriptions of their goals, methodology, background, and causal reasoning, while avoiding reliance on non-textual elements (e.g., numerical data). Because each paper centers around a single, human-authored causal graph, adheres to a standardized structure, and articulates modeling assumptions, causal graphs can be accurately annotated, and construction from the text is feasible, even when elements are not mentioned explicitly. At the same time, economics is highly cross-disciplinary, spanning many domains, which offers substantial diversity while allowing for standardized quality control. Statistics detailing the distribution of academic domains within ReCAST are presented in Appendix B.

To support large-scale collection, we utilize API access to economics papers. To ensure our corpus allows for legal redistribution of texts and diagrams, we restrict our search to the open access repositories (MDPI and PLOS). We use a keyword search for the term "causal loop diagram", a commonly used term in economics literature to describe causal graphs, resulting in 646 candidate papers. We exclude workshop papers by removing papers whose abstracts include the keywords "workshop" or "group model build", as these texts do not include sufficient details about the graph to make it identifiable. We then manually exclude all papers with graphs unsuitable for annotation, including ones with no or multiple unrelated causal graphs, purely illustrative or non-causal diagrams, and causal graphs with poor legibility or ambiguous graph elements. We retain only papers where there is a single causal graph - either as the sole causal graph, or as the culmination of a sequence of graphs - so that it can serve as the ground-truth causal graph for benchmarking. During the manual filtering, the location of each primary causal graph in the paper is labeled to ensure accurate annotation (e.g., "Top diagram on page 7").

### 3.2 Annotation

As ReCAST is a text-based benchmark, it is important that ground-truth causal graphs are converted from images to a text-based representation to be used as the ground-truth answer. We find through manual evaluation that vision-based LLMs are unsuitable for annotation due to frequent hallucinations. Therefore, we employ domain experts to annotate each causal graph into a standardized format ("source_variable" -> "sink_variable"). They are provided detailed instructions for the annotation process, including which graph elements to include, a standardized output format, and step-by-step examples. To avoid potential errors, we further instruct annotators to flag any papers that cannot be unambiguously annotated given the instructions, which we manually review for suitability.

### 3.3 Post-Processing

**Graph Post-Processing** To achieve ReCAST's goal of being a realistic causal reasoning benchmark, it is essential that the ground-truth answers are accurate. While human annotators are highly accurate at transcribing causal graphs to our standardized textual format, they are prone to minute errors. To mitigate the risk of erroneous ground-truth answers, we utilize a rigorous post-processing pipeline for ground-truth graphs. We first utilize code-based approaches to identify formatting mistakes, and attempt string matching approaches to automatically correct them. Any samples that are unable to be corrected for formatting are flagged and have the formatting manually corrected.

To verify transcription quality, a second annotator re-labeled 37 randomly chosen diagrams (778 edges, 616 nodes; $\approx 15\%$ of ReCAST). We observed 22 missing and 5 spurious edges, giving edge–level precision $= 0.995$, recall $= 0.977$, $F_1 = 0.986$, and Cohen's $\kappa = 0.94$. Variable labels showed 8 auto-correctable typos (e.g., *carbondioxide*) and 4 minor prefix/suffix omissions; no major name mismatches occurred. Full per-graph statistics and the Python script used to compute them are provided in Appendix J. We explore the possibility of utilizing code-based approaches for correction of variable naming, but find that it is prone to erroneously combining distinct variables (e.g., GDP and GNP). Therefore, we utilize LLMs for automated correction, which is detailed further in Appendix M.

**Text Post-Processing** With the causal graphs finalized, it is important to ensure that the suitability of source texts. As ReCAST is a text-based benchmark, the PDF must first be converted to a textual format. As markdown is a common format for LLMs, we choose this as our target for conversion. We use the Python library PyMuPDF to extract the raw text for each paper. The output of this step contains numerous formatting errors, such as arbitrary line breaks. We utilize a multi-step LLM pipeline, as code-based approaches are infeasible given the diversity of documents, and manual testing shows current LLMs cannot perform this process accurately in one step.

We first prompt Mistral Small (Team, 2025a) to convert from the PDF text to well-structured markdown (see Appendix N). The goal of this step is to remove non-textual elements (which are impossible to accurately represent in markdown) and entirely irrelevant elements (to streamline the task and save on computational costs). Therefore, Mistral is tasked to output the markdown autoregressively while skipping over non-textual elements (such as images, charts, or other figures), in-line citations, references, publication information, and appendices. The output of this step is a well-formatted markdown version of the paper.

**Normalization Tool**

```
{ "normalizations": [
    {"start": "text to find (start)",
    "end": "text to find (end)",
    "replacement": "new text"}, ... ] }
```

However, this may include information that makes the task trivial (such as a table including all of the sample's causal relations), or it may erroneously reference removed elements (resulting in an internally inconsistent document). To correct these issues, we utilize o3-mini (OpenAI, 2025) to remove explicit references to the causal graph and correct any references to missing elements using a normalization tool. Explicit references to the graph are unrealistic (such as a table listing every connection in the graph), as the task is to create a causal graph that does not already exist. However, we are careful to not remove other information about the graph to avoid making it unidentifiable. To make these changes, we provide the model a normalization tool to minimize output tokens, and unnecessary changes, while allowing for code-based validation.

# 4 Experiments

We examine the following aspects of LLM performance on our benchmark: (i) ReCAST's ability to reveal LLM limitations across strong open- and closed-source models, including both instruction-tuned and reasoning-optimized variants, at varied realistic causal reasoning tasks; (ii) Ablation analysis to isolate sources of error, focusing on factors including the role of unobserved confounders, and varying sizes; and (iii) Performance of LLMs under name-assisted causal graph construction, where node identification is controlled for, isolating causal inference capabilities from entity extraction accuracy.

| Attributes | Number |
|---|---|
| Number of Samples | 292 |
| Average Nodes per Graph | 24.99 |
| Average Edges per Graph | 37.37 |
| Average Length of Text | 40541.15 |
| Average Degree of Confounding per Node | 0.123 |

Table 1: ReCAST statistics. The average length of ReCAST is measured in characters for consistency.

In the following subsections, we first describe our experimental setup, including models, metrics, and formatting pipelines, and then present results under both the standard and ablation settings. Additional implementation details and extended analysis can be found in the Appendix[1].

## 4.1 Evaluation Details

To verify ReCAST's effectiveness, we first select state-of-the-art LLMs of varying model sizes and types for evaluation. Specifically, as we are interested in examining the performance of LLMs for real-world causal reasoning, we aim for diverse models to evaluate. To evaluate each open-source LLMs, we employ its official prompt format for each test sample. During generation, we use the hyperparameters recommended by its creator to ensure we utilize a strong baseline. We also set the max tokens to 100,000 to ensure that models are not penalized for utilizing long chain-of-thoughts. For closed source models, we retry on failure, to ensure that all models generate valid answers. Due to the nature of ReCAST, some samples are extremely long, and in rare cases may exceed a model's maximum context window; we do not include these failed answers to avoid artificially depressing scores. We also provide the LLM the number of expected nodes, as fixing $V$ isolates the causal graph's abstraction from the causal reasoning objective. Without doing so, there are many valid levels of granularity for the causal graph, greatly complicating automating grading, and creating a confounder for the causal reasoning task.

**Evaluated Models.** We evaluate 5 recent LLMs,

[1]Our code can be found at https://anonymous.4open.science/r/CausalBenchmark-E0C6.

varying in size and training on our task, below:

| Model | Access | Size | Type |
|---|---|---|---|
| o3-mini OpenAI | Closed | — | Reasoning |
| R1 et al. | Open | 685 | Reasoning |
| Qwen-32B Team | Open | 32 | Instruct |
| QwQ Team | Open | 32 | Reasoning |
| Llama-8B Grattafiori et al. | Open | 8 | Instruct |

Table 2: Summary of Evaluated LLMs. Size in billions of parameters.

### 4.2 Evaluation Metrics

**Traditional Metrics.** Evaluating causal graph construction from text poses unique challenges not addressed by standard metrics in causal discovery. Traditional approaches assume structured inputs and rely on exact variable and edge matches. Common metrics like Structural Hamming Distance (SHD), precision, recall, and $F_1$ require that variable names are exact matches, which will rarely hold for this task. For example, "person" and "human" are semantically nearly identical, but are not exact matches as variable names. Therefore, binary metrics are poorly suited for this task.

Given the automated evaluation aims to compare a generated graph with a ground-truth graph, Graph neural networks (GNNs) seem like a natural choice for graph similarity evaluation. However, we find that it is a poor choice for this task, due to a lack of access to the source text, a lack of fine-grained output, and lack of differentiation in performance when the task is simplified. We provide further analysis of GNN efficacy for this evaluation task, including quantitative results, in Appendix E.

**LLM-as-a-Judge Metrics.** To address these limitations, we introduce a LLM-as-a-Judge automated evaluator. It provides fine-grained feedback using the prompt and grading calculations as detailed in Appendix T and Appendix U. The prompt for the LLM judge ensures several properties are exhibited during grading. There may be varying levels of abstraction in the graph; it is important that a model is rewarded for including a relationship at all, and using the proper level of abstraction is rewarded increasingly as it gets closer to the ground-truth graph. There are many choices when making a graph that are merely stylistic; a model should not be penalized for this, even if it was different from the ground-truth (most commonly with variable naming). We utilize varying labels to allow for grading on the level of individual nodes and edges, and aggregate these metrics across the entire graph. We also aim to minimize overlap between categories, so that performance in one area does not contaminate judging in another. Additionally, it is important to use the proper level of criticism when grading; missing a core relationship to a graph or getting the causal logic wrong is a much more severe issue than choosing different variables to include on the periphery of the graph. This also allows other factors like hallucinations, reversing causality, and other factors to be appropriately incorporated into overall metrics while minimizing ambiguities in grading. The LLM judge, *R1*, is provided with the source text and ground-truth graph, and outputs a categorical assessment across several criteria for each node and edge. These are aggregated across all nodes and edges for each graph, which can be used to calculate traditional binary metrics, as detailed below.

**Precision, Recall, $\mathbf{F_1}$**: Measured on the directed edges $(u, v)$ in the predicted and reference graphs.

**Structural Hamming Distance (SHD)**: $\mathrm{SHD}(G_{\text{true}}, G_{\text{pred}}) = \#\mathrm{Adds} + \#\mathrm{Dels} + \#\mathrm{Revs}$

**Normalized SHD**: $\text{Normalized SHD} = \frac{\mathrm{SHD}}{n(n-1)}$ where $n$ is the number of unique nodes. Metrics are computed using adjacency matrices with `CausalDiscoveryToolbox`.

**Formatting and Normalization.** We enforce strict JSON formatting for graph outputs to enable automated evaluation. Since models may introduce formatting errors, we apply a post-processing step using Mistral Small (Team, 2025a) to convert malformed outputs into valid JSON while preserving the intended relationships. The correction prompt is shown in Appendix Q.

## 5 Experimental Results

### 5.1 Main Results

Table 3 presents the accuracy of all selected LLMs on the realistic causal graph generation from text task. We provide a detailed metrics of performance, including precision and recall, F1, SHD, and Normalized SHD, with corresponding standard deviations. We report for nodes and edges as applicable.

The first observation we have from Table 3 is that all evaluated models show relatively poor performance, with the best-performing model, `R1`, achieving just an average $F_1$ of 0.477 across all samples, falling far short of ideal performance. However,

| Model | Node Precision (↑) | Node Recall (↑) | Edge Precision (↑) | Edge Recall (↑) | F1 (↑) | SHD (↓) | Normalized SHD (↓) |
|---|---|---|---|---|---|---|---|
| QwQ | $0.881 \pm 0.119$ | $0.488 \pm 0.216$ | $0.802 \pm 0.201$ | $0.242 \pm 0.200$ | $0.450 \pm 0.193$ | $\underline{36.860 \pm 21.793}$ | $0.107 \pm 0.084$ |
| Llama-8B | $0.827 \pm 0.191$ | $0.359 \pm 0.244$ | $0.677 \pm 0.265$ | $0.125 \pm 0.160$ | $0.302 \pm 0.201$ | $41.197 \pm 23.631$ | $0.120 \pm 0.102$ |
| Qwen-32B | $0.862 \pm 0.117$ | $0.434 \pm 0.242$ | $0.747 \pm 0.231$ | $0.181 \pm 0.170$ | $0.381 \pm 0.200$ | $40.602 \pm 23.886$ | $0.112 \pm 0.089$ |
| o3-mini | $0.862 \pm 0.134$ | $0.459 \pm 0.231$ | $0.806 \pm 0.197$ | $0.208 \pm 0.185$ | $0.415 \pm 0.200$ | $38.481 \pm 22.120$ | $0.107 \pm 0.090$ |
| R1 | $\underline{0.893 \pm 0.108}$ | $\underline{0.522 \pm 0.217}$ | $\underline{0.817 \pm 0.183}$ | $\underline{0.260 \pm 0.196}$ | $\underline{0.477 \pm 0.187}$ | $38.193 \pm 22.491$ | $\underline{0.105 \pm 0.097}$ |

Table 3: Comparison of different models' performance on ReCAST (mean ± standard deviation). Best is underlined.

despite poor overall performance, models vary substantially. `Llama-8B` performs noticeably worse than all other models across all metrics. Comparing across metrics, models perform in roughly the same ranking for each, showing broad agreement for overall performance.

We note that the standard deviation is high across all metrics. We attribute this variable performance primarily to the realistic nature of the benchmark. Sample difficulty will vary based on factors including degree of confounding, number of nodes, number of edges, text size, and domains mimicking the diverse conditions for real-world causal reasoning. All models also exhibit significantly lower recall than precision, showing that models have an easier time generating nodes and edges that are correct, but not necessarily the same as the ground-truth graph. Raw SHD scores are high, as expected, due to the large size of graphs. However, normalized SHD is low, due to most graphs being sparse.

We also find that there is a relationship between model size and performance. The worst-performing model, `Llama-8B`, is just 8 billion parameters, while the best performing model is `R1`, with 685 billion. There is also a relationship between model performance and model type, as reasoning models consistently outperform instruction models. For example, `QwQ` is a reasoning-trained version of `Qwen-32B` with the same underlying model architecture, and outperforms the instruct model on every metric.

## 5.2 Evaluation under different Degrees of Confounding

We investigate whether poor overall performance is due to succeeding on easier samples while struggling on harder samples. To do so, we analyze the degree of confounding, which shows how much information a source text contains about its graph. To measure this, we use the following equation, which is calculated as the percentage of nodes in a graph that are not explicitly mentioned in a text.

$$\mathrm{DC} = \frac{1}{|V|} \sum_{v \in V} \begin{cases} 1, & \text{if } v \notin E, \\ 0, & \text{if } v \in E \end{cases}$$

Degree of Confounding is an important attribute of samples, as it is a natural measure of difficulty, as nodes that are explicitly mentioned are easier to identify than entirely unobserved confounders. We use *R1* to label the degree of confounding for each node in each sample, providing it the ground-truth graph and source text as a reference. Each node is classified as explicit (the node or a synonym is explicitly described in the text) or not (the node appears implicitly or indirectly, or does not appear in the text). We calculate degree of confounding for each sample using the formula detailed above, and average over all samples.

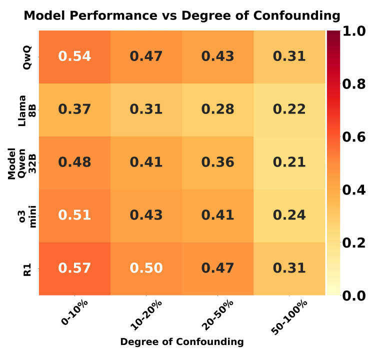


Figure 2: Heatmap of the average model scores across confounding bins (lenient definition), showcasing how degree of confounding has a large impact on model performance. It underscores LLMs' struggles to infer causal relationships when explicit references are sparse.

Using this information, the samples are categorized into four categories, ensuring that each has sufficient samples to show the effect of confounding. We explicitly do not split into evenly sized regions, as samples with extremely high degrees of confounding are unrealistic, and hence do not appear in ReCAST. As shown in Figure 2, there is a strong trend between a sample's degree of confounding and model performance (reported using $\mathrm{F}_1$). For example, the lowest-confounding bin (<10%) the best model (*R1*) achieves an average

| Model | Precision (↑) | Recall (↑) | $F_1$ (↑) | SHD (↓) | Normalized SHD (↓) |
|---|---|---|---|---|---|
| R1 | $\underline{0.513 \pm 0.202}$ | $\underline{0.512 \pm 0.072}$ | $\underline{0.502 \pm 0.146}$ | $40.558 \pm 35.173$ | $\underline{0.085 \pm 0.066}$ |
| Llama-8B | $0.219 \pm 0.150$ | $0.409 \pm 0.054$ | $0.267 \pm 0.136$ | $58.561 \pm 50.222$ | $0.142 \pm 0.110$ |
| o3-mini | $0.455 \pm 0.203$ | $0.491 \pm 0.072$ | $0.459 \pm 0.153$ | $40.654 \pm 37.081$ | $0.086 \pm 0.069$ |
| Qwen-32B | $0.290 \pm 0.171$ | $0.435 \pm 0.063$ | $0.332 \pm 0.145$ | $50.952 \pm 36.189$ | $0.123 \pm 0.085$ |
| QwQ | $0.485 \pm 0.197$ | $0.502 \pm 0.071$ | $0.483 \pm 0.143$ | $\underline{40.290 \pm 32.571}$ | $\underline{0.085 \pm 0.063}$ |
| Random | $0.131 \pm 0.101$ | $0.379 \pm 0.040$ | $0.179 \pm 0.099$ | $67.809 \pm 44.913$ | $0.158 \pm 0.093$ |

Table 4: Aggregate name-assisted metrics: precision, recall, $F_1$, SHD, and normalized SHD. Performance is poor despite providing node names, highlighting that reasoning is required for the task beyond node identification.

$F_1$ of 0.57, yet this reduces by almost half to 0.31 for samples where more than half of the nodes are confounding. This trend shows that while models struggle for all degrees of confounding, they especially struggle when they must use causal reasoning to infer information from the text. This finding is corroborated by analysis using an alternate, stricter definition of confounding (see Appendix F).

### 5.3 Evaluation under Varying Input Sizes

We also investigate the effect of input size on performance. This includes the length of the source text, number of edges in the ground-truth graph, and number of nodes in the ground-truth graph. We report the difference in performance between the small and large samples in Table 5. We measure length using number of characters rather than tokens to standardize comparison. Counterintuitively, there is a positive trend between increasing size and model performance. We investigate this by analyzing the relationship between degree of confounding and size. Due to the large variation in size, we analyze using a log scale. We find that text size, number of nodes, and number of edges each have a small, negative correlation with degree of confounding (with $R^2$ values of 0.187, 0.002, and 0.006 respectively). This explains some of this trend given the strong relationship between degree of confounding and performance. We include charts visualizing this relationship in Appendix H.

| Model | Text | Edges | Nodes |
|---|---|---|---|
| QwQ | $+0.016 \pm 0.284$ | $+0.004 \pm 0.266$ | $+0.055 \pm 0.266$ |
| Llama-8B | $-0.013 \pm 0.283$ | $+0.010 \pm 0.281$ | $-0.000 \pm 0.266$ |
| Qwen-32B | $+0.058 \pm 0.287$ | $-0.129 \pm 0.293$ | $-0.088 \pm 0.287$ |
| o3-mini | $+0.031 \pm 0.285$ | $-0.061 \pm 0.298$ | $+0.055 \pm 0.283$ |
| R1 | $+0.037 \pm 0.250$ | $+0.003 \pm 0.290$ | $+0.016 \pm 0.254$ |

Table 5: Difference in model performance (mean $\pm$ std) between the top and bottom quartiles of text length, edge count, and node count. Positive values indicate improved performance on larger instances.

### 5.4 Ablation Study: Name-Assisted Generation

To investigate the extent to which poor model performance is explained by node identification or deeper causal reasoning, we conduct an ablation experiment in which models are explicitly provided with the complete set of ground-truth node names. Under these conditions, evaluation is deterministic, isolating the effects of LLM-based grading and node identification from that of edge causality.

Table 4 shows results from this controlled scenario. While models trivially achieved perfect node-level precision and recall by design, the expected improvements in causal inference were limited: edge-level $F_1$ slightly improved for stronger models such as `R1` (by only $+0.025$) and `o3-mini` (by $+0.044$). Meanwhile, weaker models like `Qwen-32B` and `Llama-8B` saw reductions in performance, declining by $-0.049$ and $-0.035$ respectively. SHD similarly showed minor improvements, suggesting that explicitly provided node schemas do not have a consistent effect on performance.

Though the deterministic evaluation and LLM-as-a-Judge approaches are not directly comparable—due to the differing tasks and the fine-grained vs binary assessments of each measure—both evaluation approaches show similar results. Despite removing entity recognition uncertainty, models continue to perform poorly at correctly inferring causal relationships, with causal reasoning as a fundamental bottleneck. The consistent findings between both evaluation methods lends credibility to the effectiveness of the LLM judge. This ablation highlights that the poor performance of LLMs on ReCAST is due to fundamental limitations in their causal reasoning capabilities rather than being limited to errors in surface-level entity recognition.

### 5.5 Case Study - R1 Model Answer

To illustrate model performance on realistic causal reasoning, we show a benchmark sample about livelihood efficiency in the Qinba Mountains. We

**R1 Chain-of-Thought Excerpt**

. . . In the Materials and Methods, they mention. . . auxiliary variables (savings, non-farm income, per capita arable land), and constants (region area, **precipitation**, etc.). . .
County's arable land area (18) -> Cash crop cultivation (10) (if less land, shift to cash crops). . . Cash crop cultivation (10) -> Farmers' income (maybe part of non-farm income, but non-farm income is separate. Cash crop would be part of farm income. . .
county financial capital increases lead to infrastructure and loan access. . .
- Loan usage flexibility -> Diversified investments (cash crops, etc.)

Figure 3: Verbatim excerpts from reasoning trace. Ellipses added for readability and bold for emphasis.

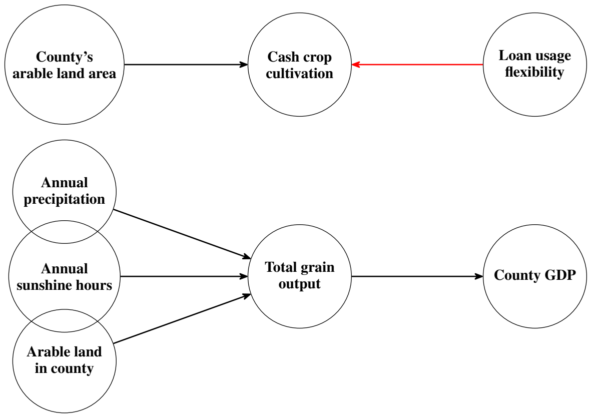


Figure 4: Top: R1 model subgraph, with the spurious link in red. Bottom: the ground-truth causal subgraph.

focus on a subset of the graph that shows the causal relationship between various climate and land factors with grain production and overall economic output. We analyze the performance of R1 on this representative subgraph. The text describes these relationships most clearly in this passage:

> Land and climate are the fundamental conditions for agricultural production. . . . therefore, sunshine, precipitation, and arable land area were selected to represent the natural capital of the county.

Even for this straightforward description, *R1* fails to correctly include this relationship in its answer. As shown in bold in Figure 3, while *R1* initially correctly identifies that precipitation is a relevant factor, it thereafter overlooks inclusion of every climate driver in its final causal graph. It does not include *Annual precipitation* or *Annual sunshine hours* despite their explicit mention as fundamental drivers. Similarly, it mentions how cash crop cultivation is a part of farm income, which is similar to GDP, but fails to elaborate on this. It also swaps the broader variable *Total grain output* for a narrow focus on *Cash crop cultivation*. It generates an erroneous link between loan usage flexibility and cash crop cultivation; however, loan flexibility relates to financial capital and coordination mechanisms, not crop selection decisions.

As shown in Figure 4, of the four ground-truth edges relevant to its subgraph, the model captures only one relationship partially by oversimplifying, omits three entirely, while adding one another erroneously. This shows how, even with straightforward textual cues, evaluated LLMs still struggle at causal reasoning. We include a complementary human expert case study for this sample in Appendix D.

## 6 Discussion and Conclusion

Understanding causal relationships from text, a core aspect of human cognition, is essential for advancing LLMs towards artificial general intelligence. However, evaluating this skill under real-world conditions has been limited by the lack of appropriate benchmarks. This paper introduced ReCAST, which is, to our knowledge, the first benchmark to assess LLM causal reasoning capabilities from text under realistic conditions. ReCAST draws diverse samples from academic literature, featuring texts varied in length, relational complexity, and domain, thereby incorporating the challenges of real-world causal reasoning.

Extensive experiments utilizing ReCAST revealed that state-of-the-art LLMs struggle significantly with this task, with the best-performing model achieving an average F1 score of just 0.477. In-depth analysis identified common pitfalls, including challenges with implicitly stated information and distinguishing relevant causal factors from surrounding contextual details across lengthy passages. Notably, performance remained poor even under low degrees of confounding or when node identification was simplified, indicating the primary bottleneck lies in causal reasoning itself. ReCAST offers a robust platform for systematically characterizing these deficiencies. It serves as a valuable tool for the research community, enabling more targeted investigations into LLM limitations. These insights directly inform the development of next-generation models, advancing efforts towards more sophisticated causal reasoning abilities in LLMs.

## 7 Limitations

While ReCAST establishes a valuable methodology for assessing realistic causal graph generation from text, several limitations should be noted. Firstly, the ground-truth causal graphs are derived from existing economics literature. While these sources are peer-reviewed and provide authentic complexity, they may reflect domain-specific simplifications or particular theoretical assumptions inherent to the original research. This establishes a practical upper bound on achievable performance by any model and means the evaluation is against one specific interpretation of causality within that text. Secondly, the use of open-access source documents, essential for benchmark accessibility, introduces the possibility that some texts were part of the evaluated LLMs' pre-training data. Our analysis regarding knowledge cutoffs (Appendix K) suggests this did not significantly impact results for the models tested, but it remains a general consideration as the precise contents of large pre-training corpora are often opaque.

Thirdly, while benchmark samples are drawn from diverse sub-domains within economics, the benchmark's overall focus on this single academic discipline may still limit the diversity of causal phenomena and linguistic expressions of causality encountered. This could influence the direct generalizability of findings to fields with different conventions. Fourthly, our methodology incorporates LLMs for several post-processing steps and, crucially, for the LLM-as-a-Judge evaluation. While efforts were made to ensure robustness (e.g., through specific prompting, tool use, and ablations with deterministic evaluation), this introduces the potential for LLM-specific biases or inaccuracies in these automated steps. For instance, the LLM Judge's assessment, despite careful calibration, might subtly favor certain graph characteristics. Finally, the task is simplified by providing models with the target number of nodes for the causal graph, which differs from the fully unconstrained nature of some real-world causal discovery scenarios where identifying the relevant variables and their appropriate level of abstraction is a key part of the challenge.

## 8 Ethics Statement

**Annotator Data** Ground-truth annotations for ReCAST were sourced from economics students participating for opt-in extra credit. No personally identifiable information (PII) besides name was collected (to assign credit), which has since been destroyed, guaranteeing complete anonymization. Participants were explicitly informed that their anonymized annotation data would be used for research use. Given this de-identification, formal IRB review was not required.

**Use of AI Assistants** AI assistants aided in literature review, drafting, coding, and data processing, under human direction and review.

**Benchmark Responsible Use** ReCAST consists of English-only academic texts (primarily in economics), which may introduce linguistic or domain-specific biases. The ReCAST dataset, consistent with its source data, is intended for research. While advancements in AI causal reasoning could be misused, ReCAST is intended to foster responsible research and development of beneficial AI systems. We are committed to transparency and will publicly release ReCAST and associated code. Source academic papers are used under their original open-access licenses (details with dataset release).

## References


T. Bratanic. 2024. Building knowledge graphs with llm graph transformer: A deep dive into langchain's implementation of graph construction with llms. Towards Data Science.

DeepSeek-AI et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuval Noah Harari. 2014. *Sapiens: A Brief History of Humankind*. Random House, New York.

Niyousha Hosseinichimeh, Aritra Majumdar, Ross Williams, and Navid Ghaffarzadegan. 2024. From text to map: a system dynamics bot for constructing causal loop diagrams. *System Dynamics Review*, 40(3):e1782.

Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. G. Adauto, M. Kleiman-Weiner, M. Sachan, and B. Schölkopf. 2024. CLadder: Assessing causal reasoning in language models. https://arxiv.org/abs/2312.04350.

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. 2024a. Cold: Causal reasoning in closed daily activities. *Preprint*, arXiv:2411.19500.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024b. Llms are prone to fallacies in causal inference. *Preprint*, arXiv:2406.12158.

Hyunjung Kim and David F. Andersen. 2012. Building confidence in causal maps generated from purposive text data: mapping transcripts of the federal reserve. *System Dynamics Review*, 28(4):311–328.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Preprint*, arXiv:2305.00050.

Uriel Anderson Lasheras and Vladia Pinheiro. 2025. CaLQuest.PT: Towards the collection and evaluation of natural causal ladder questions in Portuguese for AI agents. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 325–343, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. 2025. Explica: Evaluating explicit causal reasoning in large language models. *Preprint*, arXiv:2502.15487.

Sangmin S. Oh. 2025. Plausibly exogenous galore. Notion. Available at https://sangmino.notion.site/1a897b8106ca44eeaf31dcd5ae5a61b1?v=ff7dc75862c6427eb4243e91836e077e.

OpenAI. 2025. Openai o3-mini system card. OpenAI, January 31, 2025. https://cdn.openai.com/o3-mini-system-card.pdf.

Judea Pearl. 2009. *Causality*. Cambridge University Press, Cambridge.

Mistral AI Team. 2025a. Mistral small 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models!

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

G. A. Veldhuis, D. Blok, M. H. T. de Boer, G. J. Kalkman, R. M. Bakker, and R. P. M. van Waas. 2024. From text to model: Leveraging natural language processing for system dynamics model development. System Dynamics Review, 40(3): e1780.

Duo Wang, Yuan Zuo, Fengzhi Li, and Junjie Wu. 2024. Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings. *Preprint*, arXiv:2408.14512.

Khurram Yamin, Shantanu Gupta, Gaurav R. Ghosal, Zachary C. Lipton, and Bryan Wilder. 2024. Failure modes of llms for causal reasoning on narratives. *Preprint*, arXiv:2410.23884.

S. Yu, T. Huang, M. Liu, and Z. Wang. 2023. BEAR: Revolutionizing service domain knowledge graph construction with llm. In Monti, F., Rinderle-Ma, S., Ruiz Cortés, A., Zheng, Z., and Mecella, M. (Eds.), Service-Oriented Computing. ICSOC 2023. Lecture Notes in Computer Science, vol 14419, pp. 339–346. Springer, Cham.

B. Zhang and H. Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. arXiv preprint arXiv:2404.03868.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. Causalbench: A comprehensive benchmark for causal learning capability of llms. *Preprint*, arXiv:2404.06349.

## A Detailed Comparison with Causal Reasoning Benchmarks

This section provides a comparative overview of ReCAST against existing benchmarks for causal reasoning. While each benchmark contributes to understanding LLM capabilities, ReCAST is specifically designed to test the construction of complex causal graphs from long-form, real-world academic texts. We evaluate benchmarks based on their diversity and complexity. In this context, we consider a benchmark diverse when it spans many domains or sub-domains, or draws data from many different types of sources. Meanwhile, we deem a sample complex when there are varied degrees of confounding in samples; that is, that the benchmark features many samples that are not highly explicit, requiring that causal reasoning, rather than mere reading comprehension, to be used. Lastly, we also examine the realism (of the source data) - is it drawn from the real-world? The following table offers a high-level visual comparison, with further details on each benchmark discussed subsequently.

### Discussions of Selected Benchmarks

The following discussions provide context for the data presented in Table 6, highlighting their approaches and how we notably differ from this prior work. While these works are valuable, prior work fails to measure the causal reasoning abilities of LLMs from text under real-world settings. We highlight the approach of each benchmark, and how they compare to ReCAST.

Several benchmarks concentrate on pairwise causal relations, include inputs that are highly explicit, or have inputs that are generated synthetically or are crafted as short texts by hand. We differ from these approaches by aiming to construct large graphs from real-world literature. **ExpliCa (Miliani et al., 2025)** examines how LLMs understand explicit connectives in sentence pairs, resulting in 2-node links. By design, it uses very short, often crafted inputs and focuses on explicit cues, thereby avoiding the complexities of implicit causality and information integration from extensive texts that ReCAST targets. Meanwhile, **LLM Fallacies (Joshi et al., 2024b)** employs short, synthetic scenarios to test pairwise causal inference, focusing on logical fallacies when LLMs are presented with explicit non-causal information. While this addresses a specific type of reasoning complexity, its synthetic and brief inputs differ greatly from the real-world, extensive texts and broader graph construction task in ReCAST. **Plausibly Exogenous Galore (Oh, 2025)** serves as an interesting bridge, as it uses long economics documents similar to ReCAST. However, its task is to find only the main pairwise link for the entire document, which greatly limits its diversity and complexity.

Another group of benchmarks attempt graph construction, but typically rely on short, simplified, or synthetic descriptions. These lack the depth and realism of long, real-world texts. **From text to map (Hosseinichimeh et al., 2024)** generate relatively small graphs (max 9-15 nodes) from concise, hand-crafted descriptions. Such inputs inherently limit textual diversity and likely feature more explicit causal links, sidestepping the challenge of parsing lengthy, nuanced documents with varying levels of confounding. **Failure Modes of LLMs for Causal Reasoning on Narratives (Yamin et al., 2024)** also uses short, often synthetic or CauseNet-derived narratives for constructing linear chain graphs (max 20 nodes). While it explores LLM biases and indirect effects, its input lacks the textual diversity and structural graph complexity of ReCAST, and its narratives are purpose-built rather than reflecting the reasoning of real-world conditions.

Other benchmarks focus on a sentence-level analysis, or offer broader surveys of causal tasks where individual components may use non-primary inputs or address different facets of reasoning. **From Text to Model (Veldhuis et al., 2024)** (NLP for SD) measures the ability of LLMs to classify individual sentences from real-world texts for causality. While it uses real-world text, it uses small excerpts, and avoids the complexities of reasoning over large documents. **Causal Reasoning Survey (Kıcıman et al., 2024)** provides a wide-ranging overview of LLM capabilities across multiple causal tasks. However, its sub-tasks often use short or structured inputs (e.g., variable lists for graph construction, concise vignettes for reasoning), which differ from ReCAST's reliance on extensive, unmodified academic texts for end-to-end graph extraction.

To the best of our knowledge, ReCAST is the first benchmark for LLMs to measure the causal reasoning abilities from diverse, complex, real-world texts. While previous benchmarking efforts have explored lengthy, diverse, and complex texts, causal reasoning, and real-world conditions separately, we are the first to do so at once.

| Benchmark | Primary Task | Long Text | Input Type | Max Nodes | Diverse | Complex | Realism |
|---|---|---|---|---|---|---|---|
| **ExpliCa** (Miliani et al., 2025) | Pairwise ID | ✗ | Sentences | 2 | ✗ | ✗ | ✗ |
| **LLM Fallacies** (Joshi et al., 2024b) | Pairwise Inference | ✗ | Scenario/Vignette | 2 | ✗ | ⚠ | ✗ |
| **Plausibly Exogenous** (Oh, 2025) | Pairwise ID | ✓ | Full Document | 2 | ✓ | ⚠ | ✓ |
| **From text to map** (Hosseinichimeh et al., 2024) | Graph Construction | ✗ | Short Narratives | 15 | ✗ | ✗ | ✓ |
| **Failure Modes** (Yamin et al., 2024) | ID/Graph Construction | ✗ | Short Narrative | 20 | ✗ | ⚠ | ⚠ |
| **From Text to Model** (Veldhuis et al., 2024) | Sentence Classification | ✗ | Sentences | N/A | ✗ | ⚠ | ✓ |
| **Causal Reasoning Survey** (Kıcıman et al., 2024) | Multiple Tasks | ✗ | Mixed | Varies | ⚠ | ⚠ | ⚠ |
| **CLadder** (Jin et al., 2024) | Causal QA | ✗ | Narrative/Question | 4 | ✗ | ✓ | ✗ |
| **CausalBench** (Zhou et al., 2024) | Causal Structure ID | ✗ | Query/Question | 109 | ⚠ | ⚠ | ✓ |
| **COLD** (Joshi et al., 2024a) | Causal QA | ✗ | Query/Question | 33 | ⚠ | ✓ | ✓ |
| **CaLQuest.PT** (Lasheras and Pinheiro, 2025) | Causal QA | ✗ | Query/Question | N/A | ✓ | ✓ | ✓ |
| **ReCAST (Ours)** | Graph Construction | ✓ | Full Document | 140 | ✓ | ✓ | ✓ |

Table 6: Visual comparison of ReCAST with other causal reasoning benchmarks. **Primary Task** (e.g., Graph Construction: Graph Construction; Pairwise ID: Pairwise Identification; QA: Question Answering). **Long Text**: Indicates if the benchmark primarily uses long textual inputs. **Input Type** (e.g., Document, Narrative, Scenarios, Queries, Sentences). **Max Nodes**: Maximum nodes per instance for graph construction or the underlying model. Symbols for realism criteria (**Diverse**, **Complex**, **Realism**): ✓: Fully meets criterion; ⚠: Partially meets criterion; ✗: Does not substantially meet criterion, relative to ReCAST's focus on long-text graph extraction.

# B Domain Statistics

| Model | 0–1 Domains | 2 Domains | 3+ Domains | Base Accuracy |
|---|---|---|---|---|
| R1 | 0.493 ± 0.176 | 0.457 ± 0.208 | 0.499 ± 0.125 | **0.477 ± 0.125** |
| QwQ | 0.461 ± 0.178 | 0.446 ± 0.210 | 0.416 ± 0.164 | 0.450 ± 0.194 |
| o3-mini | 0.409 ± 0.192 | 0.421 ± 0.213 | 0.415 ± 0.172 | 0.415 ± 0.201 |
| Qwen-32B | 0.375 ± 0.187 | 0.388 ± 0.215 | 0.364 ± 0.187 | 0.381 ± 0.201 |
| Llama-8B | 0.323 ± 0.213 | 0.290 ± 0.199 | 0.262 ± 0.140 | 0.302 ± 0.201 |

Table 7: Model performance by number of associated domains per sample, showcasing how models perform across different domain complexities, classified as 0–1, 2, or 3+ co-domains. Notably, R1 achieves the highest accuracy in single-domain samples, suggesting that models may struggle with cross-domain reasoning. Conversely, QwQ and o3-mini perform slightly better with increased domain diversity, reflecting some adaptability to multi-domain causal reasoning.

To be able to assess the domain coverage of ReCAST, we first manually create broad domain categories, aiming to minimize overlap between domains while simultaneously ensuring that all samples belong to at least one domain. It is valuable to go beyond binary classification (economics or not), as information about domains in the samples is valuable for analysis of model performance. We manually prompt engineer to create the domain categories, aiming to minimize overlap between domains and ensure that each sample falls into at least one domain. To generate initial domain categories before manual optimization, we prompt LLMs to generate categories based on all paper titles. To minimize the risk of false positives, we aim for conservative domain labeling. For example, we choose the name "Environmental & Earth Sciences" for domain #5 to ensure that economics papers about sustainability are correctly classified into domains #2 and #5. We use *R1* model to automatically classify each sample by domain, as manual validation showed high classification accuracy and adherence to output formatting. Each paper's title and abstract as context to ensure sufficient context to classify while minimizing total tokens. We use the following prompt:

**Domain Classification Prompt**

```
You are an expert at correctly labeling
domains. You will be given a published
paper's title and abstract. You will label
each other domain of the paper based on the
content. You may pick more than one domain
when applicable. Only choose domains from
the list below. Choose every domain that
is present in the paper. There may not
be any domains from the options present
in the paper. Respond with the numbers
corresponding to the domains in a JSON
list with no other text.

The domains are:
1: Agriculture & Food Systems
2: Economics & Public Policy
3: Education
4: Engineering & Technology
5: Environmental & Earth Sciences
6: Medicine

EXAMPLE INPUT:
Title: "The impact of AI on job markets"
Abstract: "This paper explores the impact
of AI on job markets and the future of
work."

EXAMPLE OUTPUT:
{ "domains": [2, 4] }
```

Each sample that it classified as non-economics is manually reviewed for accuracy, with correctly classified non-economics samples excluded from the benchmark. We manually review each sample classified as non-economics for accuracy to ensure valid benchmark samples are not erroneously excluded. After excluding non-economics papers, we finalize the benchmark samples, resulting in 292 included samples. Figure 5 shows the distribution of the number of non-economics domains for each sample, with an average of 1.67 non-economics domains per sample. We note that it is expected that almost no samples had no non-economics domains, as economics is inherently multidisciplinary.

As previously described, each sample in ReCAST belongs to Economics and at least one additional domain. On average, each sample spans 2.35 domains, with Engineering & Technology and Environmental & Earth Sciences being the most prevalent.

Performance is largely stable across samples with 0-1, 2, or 3+ co-domains. Interestingly, `R1` scores highest on single-domain samples, suggesting less cross-domain complexity may support stronger graph alignment. In contrast, `QwQ` and `o3-mini` slightly improve with increased domain count, showing some benefit from interdisciplinary contexts.

Overall, these results emphasize the importance of domain-sensitive evaluation. Variation in model performance across domains confirms that ReCAST captures diverse economic sub-domains in a way that differentiates LLM capabilities under realistic conditions.

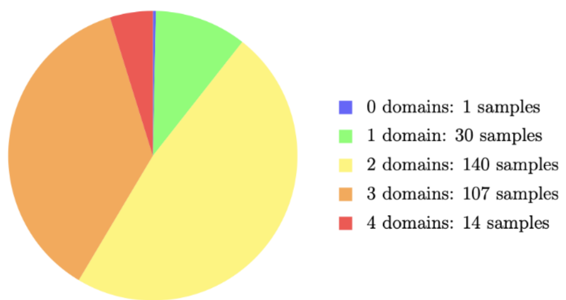


Figure 5: Distribution of domain counts per sample ($n = 292$), illustrating the multidisciplinary nature of economics-based samples in ReCAST. Most samples are associated with multiple domains, highlighting the inherent complexity of real-world economic causal reasoning.

While domain does have an effect on performance, the magnitude of such effect is not large. This is not surprising, as LLMs have seen these domains many times before during training. As shown in Figure 6, R1 (et al., 2025) boasts the highest performance across all domains, except for

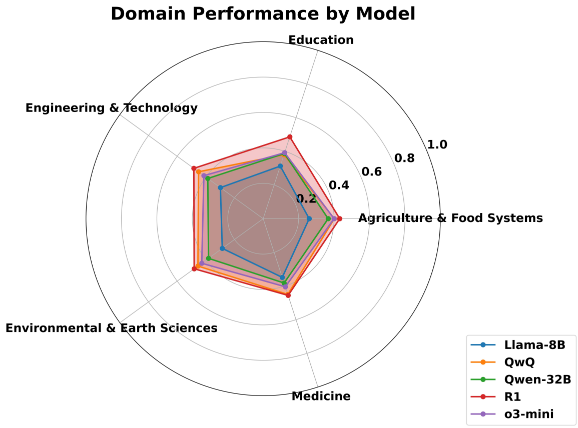


Figure 6: Radar chart of domain-specific accuracies for each model, depicting the accuracy of different models across specific domains. Variability in performance across domains highlights the contextual challenges faced by models in domain-specific causal inference, pointing towards the need for domain-sensitive model adjustments.

Agriculture & Food Systems, where QwQ (Team, 2025b) narrowly outperforms it. However, all models perform poorly overall regardless of domain, showing that this is not the factor that leads to poor causal reasoning.

Figure 7 visualizes how often each domain appears in the benchmark samples. Engineering & Technology and Environmental & Earth Sciences comprise the majority of samples, appearing in over 60% of samples. As these domains appear more frequently for PLOS and MDPI economics papers, it is not surprising that they compose an outsized portion of the dataset. Additionally, we intentionally choose broad categories for domains to ensure sufficient samples for analysis, as there is substantial diversity within each domain. Even for smaller domains, there are sufficient samples to analyze performance.

We also include the distributions of journals the samples were originally published in as an alternative measure of domain. As we intentionally choose broad domain categories, this can act as a measure of sub-domain.

# C Effect of Length of Reasoning Trace

We investigate how the length of chain-of-thought reasoning affects model performance. For each open-source reasoning model, we split data into one thousand token wide bins, and display the quantities of each token amount in Figure 9 and Figure 10. Interestingly, QwQ has some reasoning

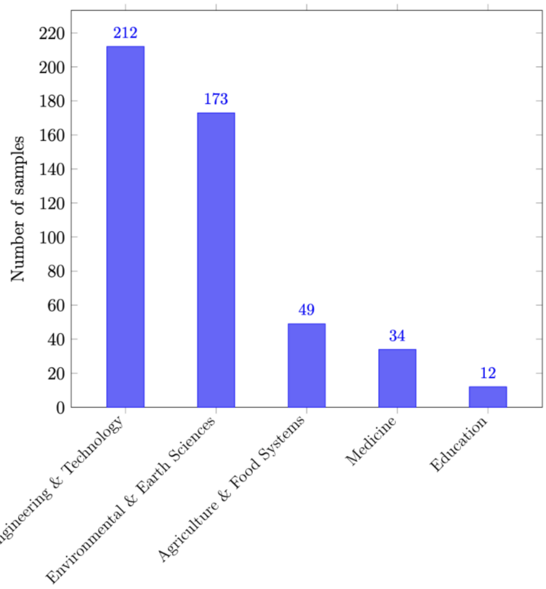


Figure 7: Domain distribution across 292 samples with processed text (excluding economics).

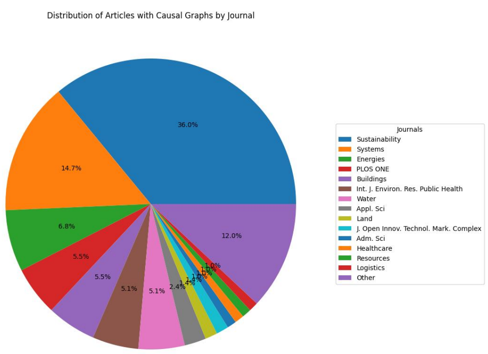


Figure 8: Distribution of journal for all ReCAST benchmark samples.

traces which are far longer than the longest reasoning traces from R1, which we attribute to the different training for each of these models. Additionally, manual inspection showed that some of these longer traces were due to repeatedly making small changes to formatting, which indicates that these responses did not spend more time on actual reasoning, and may have largely been due to QwQ's worse performance at formatting.

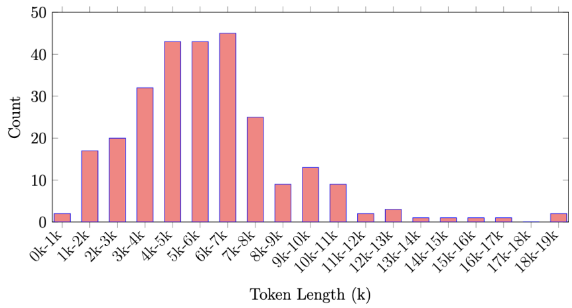


Figure 9: QwQ token length distribution. The data suggests that QwQ maintains a relatively consistent distribution across token lengths, reflecting its structured handling of reasoning chains. However, performance did not scale proportionally with longer traces, indicating limitations in handling extended reasoning efficiently.

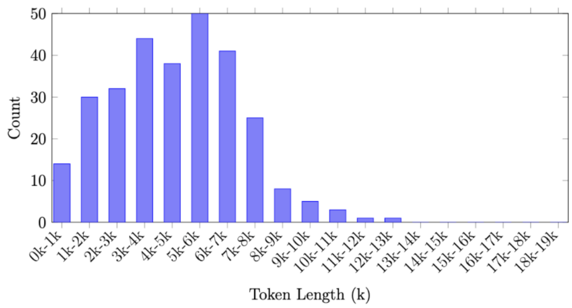


Figure 10: Distribution of reasoning trace lengths (in tokens) for R1 across benchmark samples. The model tends to produce mid-length reasoning traces (2,000–7,000 tokens), with very few exceeding 12,000 tokens.

## D Human Domain Expert Case Study

While large language models struggle to construct accurate causal graphs under realistic conditions, this task is feasible for human experts. To demonstrate this, we conduct a small-scale case study where an expert economist constructs a causal graph from the same input text used in the R1 model case study (subsection 5.5). Due to the highly time-intensive nature of human causal graph construction and annotation (Kim and Andersen, 2012), we include only this case study for this sample. We show only a representative subgraph slice of their answer to allow for direct comparison, as full graphs are too large to allow for concise comparisons. As shown in Figure 11, the expert successfully captures the three core causal drivers: `annual precipitation`, `annual sunshine hours`, and `County arable land area`. These are all mapped to a unified economic outcome, `total household income`, reflecting a broader—but valid—abstraction of the two-hop causal path present in the ground-truth graph (`Total grain output` → `County GDP`). This demonstrates that the task is tractable for humans, and the importance of evaluation that differentiates stylistic variations from mistakes in the core causal logic.

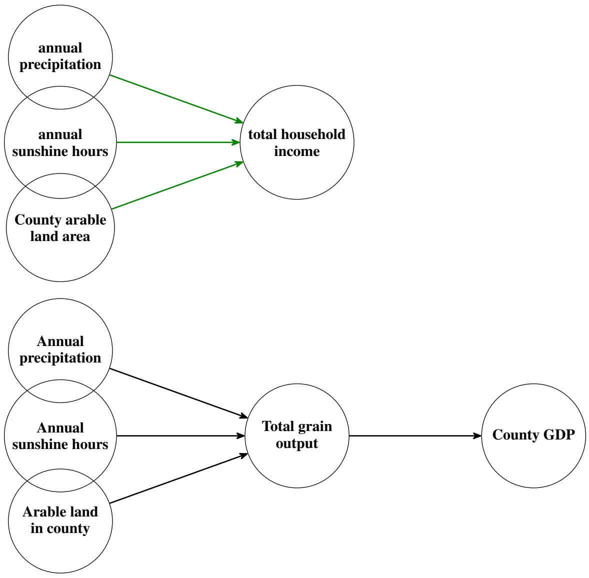


Figure 11: Top: Subgraph generated by a domain expert, with node labels shown verbatim. Bottom: Ground-truth subgraph. The expert's graph aggregates intermediate steps (e.g., `Total grain output` and `County GDP`) into `total household income`, but preserves all core causal relationships. This highlights expert-level ability to abstract while retaining semantic fidelity.

## E Efficacy of GNNs

As we aim for an automated graph-based metric, a graph neural network (GNN) is a natural first choice. However, they have several flaws for acting as an evaluator of this task. Methods like Token Embedding-Aligned Graph Language Model (TEA-GLM) (Wang et al., 2024) produce embeddings for graphs, allowing similarity to be measured via cosine distance. However, these approaches fall short in settings like ours that require semantic fidelity and textual grounding. First, GNN-based methods operate purely over graph structure and do not have access to the source text, making them unable to evaluate whether a predicted graph is faithful to the information provided. Second, they reduce a graph comparison to a single scalar score, such as cosine similarity, which offers little interpretability and no insight into specific errors in nodes or edges. Third, we find in practice that GNN embeddings are insensitive to meaningful differences: in our ablation (see Table 8), models that were explicitly given the correct node names showed nearly identical scores to those that were not, highlighting their lack of resolution. As such, while GNN-based methods remain a compelling direction for graph-level embedding, we find them unsuitable for evaluating text-grounded causal graphs where variable naming, semantic meaning, and abstraction play a critical role.

| Model | Standard | Name-Assist | $\Delta$ |
|---|---|---|---|
| `R1` | 0.926 | 0.924 | $-0.002$ |
| `Llama-8b` | 0.909 | 0.910 | $+0.001$ |
| `o3-mini` | 0.922 | 0.930 | $+0.008$ |
| `Qwen-32b` | 0.912 | 0.917 | $+0.005$ |
| `QwQ` | 0.922 | 0.923 | $+0.001$ |

Table 8: Cosine similarity ($\Delta$ = Name-Assist − Standard) under standard vs name-assisted TEA-GLM conditions. Minimal differences indicate that Name-Assist does not significantly enhance alignment with ground truth. The results raise questions about the effectiveness of name-assisted strategies for improving causal graph accuracy.

Table 8 shows the mean cosine similarity between the TEA-GLM embedding of each generated graph and its ground-truth counterpart under both conditions. In addition to the previously identified flaws, these results cast doubt onto the feasibility of GNNs as evaluators for this task. As the table shows, the maximum increase when the model is given the ground-truth node names (prompt described in Appendix R) is only +0.008 (for `o3-mini`), and one model (`R1`) even decreases by –0.002. These negligible differences cast doubt on the evaluation capabilities of the graph embedding model for this task, as substantial information being provided to models has little effect on the final embedding score.

# F Alternative Measure of Degree of Confounding

As shown in Figure 2, degree of confounding has a noticeable effect on model performance. For this, we determine whether a node is confounding by determining if it is explicitly mentioned in the text (or not). Given the large effect of degree of confounding in performance, we explore if this relationship holds under different measures of the degree of confounding. As detailed in Appendix L, each node in each graph is labeled as either explicit (the node was explicitly mentioned in the text), implicit (the node was implicitly or indirectly mentioned), or absent (the node was entirely unmentioned). Previously, we showed performance for the "lenient" measure of confounding, where we measure whether a node was mentioned explicitly, or if it was implicit/absent. We recalculate the degree of confounding for each sample for the "strict" measure as shown below.

$$\mathrm{DC}_{\mathrm{strict}} = \frac{1}{|V|} \sum_{v \in V} \begin{cases} 1, & \text{if } v \in A, \\ 0, & \text{if } v \in E \cup I \end{cases}$$

Figure 12: Definition of $\mathrm{DC}_{\mathrm{strict}}$ (degree of confounding) under the strict criterion, providing a quantitative measure of causal complexity. For each node $v$ in a sample $V$, we determine whether it is absent ($A$), or either explicit ($E$) or implicit ($I$). We count only absent nodes toward the confounding score and compute the average over all nodes in the sample.

Under this alternative measure of degree of confounding, at a given level of confounding, the "strict" measure is expected to be more difficult, as a node that is entirely absent is harder to identify than one that is implicitly or indirectly described. As shown in Figure 13, the negative relationship between degree of confounding and performance holds. Performance is also consistently worse for all models under the strict confounding definition rather than the lenient one, as expected, adding credibility to the validity of the automated labeling.

# G Base Model Case Study

As ReCAST is designed to measure the capabilities of LLMs on this task, we investigate the performance of base models. We use the Qwen-2.5-7B base model (Team, 2024), and prompt it to generate an open-ended graph. As base models are trained

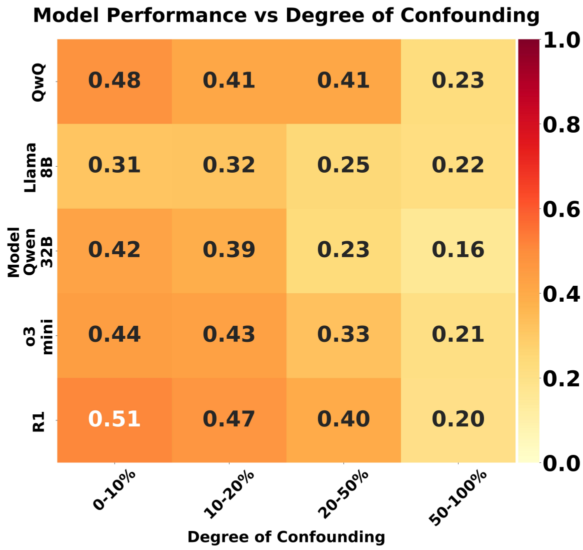


Figure 13: Average model scores across 20 % confounding bins (strict definition).

only to complete the text rather than respond to user instructions, it is unsurprising that completions were nonsensical. As with all responses, we prefill the models' response to begin with `<think>` to steer it towards using chain-of-thought thinking before responding, which in rare cases results in a valid model output.

**Endlessly Repeating Base Model Output**

```
<think>Assistant
Assistant
Assistant
Assistant
Assistant
Assistant
Assistant
```

For endlessly repeating generations, they output the same tokens until reaching the maximum tokens. For this reason, we reduce the maximum tokens for Qwen-2.5-7b to 10,000 for computational reasons, as valid generations did not come close to this maximum. Another common failure mode was immediately outputting the end of string token, resulting in no output tokens being outputted. Another failure mode was nonsensical generations, such as generating Chinese despite the text and instructions being in English. We show this example below.

**Foreign-Language Base Model Output**

````
<think>建立一个因果关系图需要遵循以下步骤：

1. 确定因果关系：首先需要确定因果关系，即哪些变量是因，哪些变量是果，以及它们之间的关系是什么。在本例中，我们已经确定了一些变量之间的因果关系，例如生产、生活和生态之间的关系。

2. 创建节点：每个变量都可以作为一个节点，用圆圈或方框表示。例如，我们可以创建三个节点，分别表示生产、生活和生态。

3. 连接节点：使用箭头连接节点，表示因果关系的方向。例如，我们可以从生产节点连接到生活节点，表示生产对生活有影响。

4. 标注节点：为每个节点添加标签，表示其含义。例如，我们可以为生产节点添加"生产"标签，为生活节点添加"生活"标签，为生态节点添加"生态"标签。

5. 添加其他变量：除了已知的变量外，还可以添加其他变量，以更全面地描述因果关系。例如，我们可以添加人口、GDP、土地利用等变量。

6. 检查因果关系：最后，需要检查因果关系是否合理，是否存在循环因果关系等。例如，生产对生活有影响，生活对生产也有影响，但这种关系是否合理需要进一步验证。

根据以上步骤，我们可以建立一个因果关系图，其中包含生产、生活、生态三个节点，以及它们之间的因果关系。具体实现代码如下：</think>```json
{
    "relationships": [
       {"source": "生产", "sink": "生活"},
       {"source": "生活", "sink": "生产"},
       {"source": "生产", "sink": "生态"},
        {"source": "生态", "sink": "生活"}
    ]
}
```
````

We list some other interesting or informative outputs from the base model as a reference.

**Another Endlessly Repeating Base Model Output**

````
<think> editText
editText
<think> editText
editText
<think> editText
editText
<think> editText
````

In rare occurrences, the model generated valid-looking responses, although the graph is far smaller than expected. We attribute its occasional success to the practice of including a small amount of instruction data into base models, which allows it to occasionally exhibit behavior normally reserved to instruction-tuned models.

**Coherent Base Model Output**

````
<think> To generate a causal graph for the provided markdown text, we first need to identify the relationships between the different factors and how they influence each other.  Based on the information provided, we can create a causal graph with the following relationships:

```json
{
    "relationships": [
      {"source": "technological breakthrough",
         "sink": "ITI development"},
        {"source": "management level",
         "sink": "ITI development"},
        {"source": "resource reserve",
         "sink": "ITI development"},
      {"source": "agglomeration spillovers",
         "sink": "ITI development"},
        {"source": "policy assistance",
         "sink": "ITI development"},
        {"source": "industrial transfer",
         "sink": "ITI development"},
        {"source": "market demand",
         "sink": "ITI development"}
    ]
}
```

This causal graph shows the different factors and their influence on the ITI development. Each node represents a factor, and each arrow represents the influence of one factor on another. The causal graph is a simple representation of the relationships between the different factors and how they impact the ITI development.
````

## H Size vs. Confounding

We provide the following charts as a reference to visualize the relationship between size and degree of confounding.
````

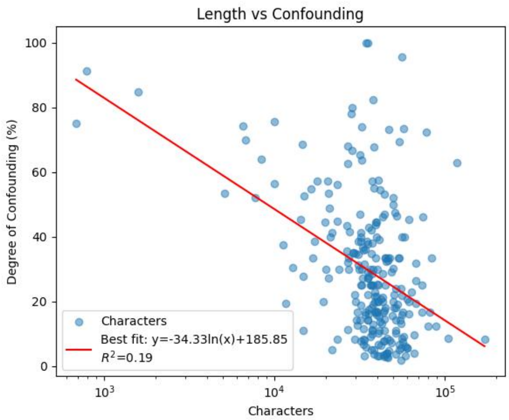


Figure 14: **Relationship between text length and degree of confounding.** This scatter plot shows each sample's character count on the x-axis versus its degree of confounding on the y-axis. There is a modest negative correlation ($R^2$ = 0.187), indicating that longer texts tend to include fewer unobserved confounders. This helps explain why models perform slightly better on larger input instances, as reduced confounding makes causal edges easier to identify.

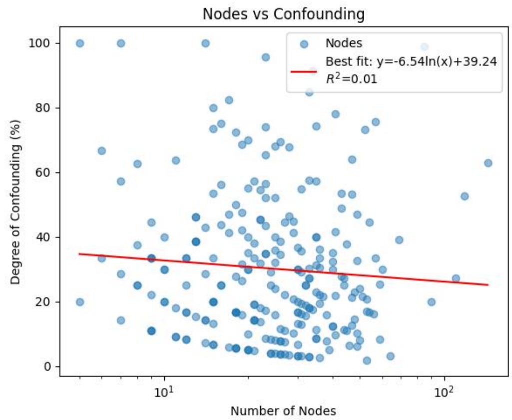


Figure 15: **Relationship between nodes and degree of confounding.** Here, each point represents a sample's ground-truth graph node count plotted against its confounding level. The near-zero correlation ($R^2$ = 0.002) shows that graph size hardly affects how many variables are unobserved in the text. This suggests that node count alone does not predict the difficulty introduced by hidden confounders.

# I Computational Costs

Despite the large size of ReCAST samples, its execution is notably quite computationally efficient. The total monetary cost for all experiments, encompassing the evaluation of all five LLMs across the main task and all ablation studies, including the LLM-as-a-Judge evaluations, remained under $250. This affordability is largely attributed to the use of prompt caching for the LLM judge. While the initial processing of the lengthy source texts incurs a significant input token cost for the judge, this cost is a one-time expense per benchmark sample. Subsequent judgments on different model outputs for the same sample, or re-evaluations, benefit greatly from caching the expensive text embedding, making the iterative evaluation process highly economical. This efficient design ensures that ReCAST can be utilized and extended by researchers without imposing prohibitive computational or financial burdens.

# J Inter-Annotator Agreement Details

To ensure the accuracy of the benchmark ground-truth graphs, we measured inter-annotator agreement by having a second annotator independently transcribe 37 randomly selected causal diagrams from the source papers ($\sim 15\%$ of the full dataset). These diagrams included 778 directed edges and 616 nodes in total. We compare the two annotators' transcriptions at both the node-level and edge-level, and compute standard metrics: precision, recall, $F_1$ score, SHD, and normalized SHD.

**Edge-Level Agreement.** Out of 778 annotated edges, 22 edges were missed by the second annotator (false negatives), and 5 extra edges were incorrectly added (false positives). There were no instances of edges that had flipped directions, nor nodes that were entirely missed. This yields:

$$\begin{aligned}
\text{Precision} &= \frac{TP}{TP+FP} = \frac{751}{751+5} = 0.993, \\
\text{Recall} &= \frac{TP}{TP+FN} = \frac{751}{751+22} = 0.972, \\
F_1 &= 0.986, \\
\text{SHD} &= 27, \\
\kappa &= 0.94
\end{aligned}$$

Cohen's $\kappa$ statistic reflects near-perfect agreement at the edge level and is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where $p_o$ is the observed agreement and $p_e$ is the expected agreement by chance, computed over all possible directed pairs.

**Node-Level Agreement.** Among the 616 nodes, we observed:

- 8 auto-correctable typos (e.g., "carbondioxide" $\rightarrow$ "carbon dioxide").

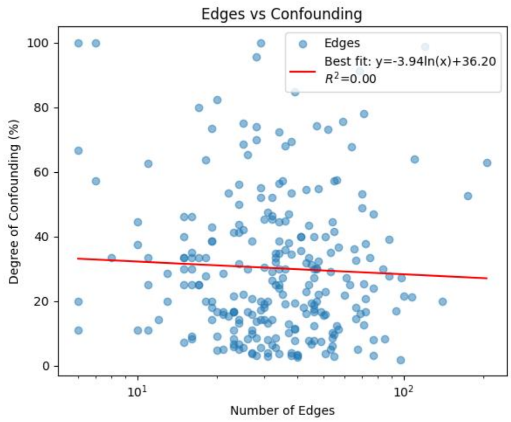


Figure 16: **Relationship between edges and degree of confounding.** This chart plots the count of true causal edges for each sample against its confounding degree. With a very small negative correlation ($R^2$ = 0.006), edge density has minimal impact on how many confounders remain implicit. Models thus face similar challenges inferring hidden variables regardless of edge complexity.

- 4 minor differences (e.g., prefix/suffix omissions such as "$CO_2$ emissions" vs. "emissions").
- No major label mismatches (0 spurious or missing nodes).

This results in a node-level $F_1$ score of 0.997.

**Sample-Level Error Overview.** We provide a high-level summary of the agreement between annotators.

The following is a summarized view of the per-sample disagreement between annotators.

This agreement analysis shows that human annotators were highly consistent, with virtually no spurious nodes and very few edge disagreements. These results validate the overall accuracy and reliability of the benchmark's gold-standard graphs.

# K Effect of Knowledge Cutoff

To investigate whether the models' performance on ReCAST is influenced by their knowledge cutoff dates (i.e., whether they might have encountered some of the source texts during their pre-training), we conducted an analysis comparing performance on papers published before and after a nominal cutoff. We selected two models for this analysis, `Llama-8B` and `o3-mini`, as their stated knowledge cutoffs are prior to 2024. This allows us to treat papers published in 2024 and onwards as more likely "unseen" by these models.

Our dataset contains 35 samples derived from papers published in 2024 or later. For `Llama-8B`, 229 samples were from papers published before 2024, and for `o3-mini`, 235 samples were from papers published before 2024. We compare the average F1 scores achieved by these models on these two subsets of ReCAST.

Table 11 shows the performance of the selected models on the 35 samples from papers published in 2024 onward. Table 12 shows their performance on samples from papers published before 2024.

Observing the mean F1 scores, both models show slightly higher average performance on the more recent papers. For `Llama-8B`, the average F1 score increased from 0.300 on older papers to 0.315 on recent papers. For `o3-mini`, the average F1 score increased from 0.410 to 0.452.

To determine if these differences are statistically significant, we performed a Mann-Whitney U test comparing the F1 scores from the "recent" (2024 onwards) and "older" (pre-2024) groups for each model.

- For **`Llama-8B`**: The p-value was 0.5356. (Recent mean F1: 0.315, Older mean F1: 0.300).
- For **`o3-mini`**: The p-value was 0.4074. (Recent mean F1: 0.452, Older mean F1: 0.410).

In both cases, the p-values are well above the common significance threshold of 0.05. This indicates that there is no statistically significant difference in performance between papers published before and after the models' nominal knowledge cutoff dates for the ReCAST task.

This finding suggests that the models' ability (or inability) to infer causal relationships from the provided texts in ReCAST is not strongly dependent on whether they might have encountered the specific source documents during pre-training. The task, by its nature, requires reasoning based on the extensive context provided within each sample, rather than direct recall of information from specific papers. The consistent performance across older and potentially "unseen" recent texts further underscores that the challenges highlighted by ReCAST are rooted in fundamental causal reasoning capabilities rather than familiarity with the source material.

| Category | Precision | Recall | $F_1$ | SHD | Norm. SHD | Cohen's $\kappa$ |
|---|---|---|---|---|---|---|
| Nodes | $0.9943 \pm 0.0204$ | $1.0000 \pm 0.0000$ | $0.9970 \pm 0.0106$ | $0.1081 \pm 0.3879$ | $0.0062 \pm 0.0223$ | N/A |
| Edges | $0.9933 \pm 0.0182$ | $0.9830 \pm 0.0509$ | $0.9876 \pm 0.0335$ | $0.7568 \pm 2.3756$ | $0.0233 \pm 0.0614$ | $0.9865 \pm 0.0367$ |
| Combined | $0.9939 \pm 0.0134$ | $0.9897 \pm 0.0329$ | $0.9916 \pm 0.0224$ | $0.8649 \pm 2.6627$ | $0.0162 \pm 0.0424$ | N/A |

Table 9: This table reports the mean and population standard deviation of key evaluation metrics over the 37 reconciled graphs. Precision, recall, and $F_1$ quantify label and edge detection accuracy. SHD is the count of false positives plus false negatives per graph, and the normalized SHD scales this by the total number of gold elements. Cohen's kappa is provided for edges only, since it relies on a clearly defined set of negative instances (all possible directed non-edges); it is not defined for node labeling or the combined set where the universe of "non-nodes" or joint negatives is ambiguous. The N/A entries indicate those cases where kappa cannot be meaningfully calculated.

| Article ID | # Nodes | # Edges | Node FP | Node FN | Edge FP | Edge FN |
|---|---|---|---|---|---|---|
| 645 | 20 | 24 | 0 | 0 | 0 | 0 |
| 630 | 25 | 32 | 0 | 0 | 0 | 2 |
| 617 | 27 | 23 | 0 | 0 | 0 | 0 |
| 588 | 10 | 16 | 0 | 0 | 0 | 0 |
| 574 | 19 | 32 | 0 | 0 | 0 | 0 |
| 566 | 24 | 23 | 1 | 0 | 0 | 1 |
| 558 | 21 | 24 | 0 | 0 | 0 | 0 |
| 552 | 16 | 15 | 0 | 0 | 0 | 0 |
| 536 | 37 | 23 | 0 | 0 | 0 | 0 |
| 497 | 15 | 20 | 0 | 0 | 1 | 0 |
| 491 | 12 | 21 | 0 | 0 | 0 | 1 |
| 486 | 28 | 32 | 0 | 0 | 0 | 0 |
| 481 | 20 | 28 | 0 | 0 | 0 | 0 |
| 458 | 18 | 24 | 0 | 0 | 0 | 0 |
| 449 | 43 | 88 | 0 | 0 | 0 | 0 |
| 440 | 12 | 15 | 1 | 0 | 0 | 0 |
| 435 | 26 | 27 | 0 | 0 | 2 | 5 |
| 410 | 9 | 16 | 0 | 0 | 0 | 0 |
| 393 | 10 | 14 | 0 | 0 | 0 | 0 |
| 362 | 11 | 16 | 0 | 0 | 0 | 0 |
| 306 | 35 | 28 | 0 | 0 | 0 | 0 |
| 303 | 9 | 10 | 0 | 0 | 0 | 0 |
| 642 | 18 | 23 | 0 | 0 | 0 | 0 |
| 259 | 19 | 48 | 2 | 0 | 1 | 12 |
| 235 | 23 | 26 | 0 | 0 | 1 | 1 |
| 200 | 16 | 24 | 0 | 0 | 0 | 0 |
| 156 | 10 | 13 | 0 | 0 | 0 | 0 |
| 95 | 26 | 39 | 0 | 0 | 0 | 0 |
| 90 | 15 | 19 | 0 | 0 | 0 | 0 |
| 74 | 13 | 22 | 0 | 0 | 0 | 0 |
| 59 | 10 | 14 | 0 | 0 | 0 | 0 |
| 43 | 15 | 17 | 0 | 0 | 0 | 0 |
| 42 | 14 | 20 | 0 | 0 | 0 | 0 |
| 589 | 12 | 16 | 0 | 0 | 0 | 0 |
| 188 | 12 | 17 | 0 | 0 | 0 | 0 |
| 168 | 15 | 19 | 0 | 0 | 1 | 0 |
| 163 | 9 | 11 | 0 | 0 | 0 | 0 |
| **TOTAL** | **674** | **879** | **4** | **0** | **5** | **22** |

Table 10: Excerpt of per-sample disagreements between two annotators. Total disagreements: 4 node FPs, 5 edge FPs, 22 edge FNs. **Inter-annotator reconciliation for 37 graphs.** *Node FP*=sum of minor+major node-label discrepancies; *Node FN*=no missing nodes observed; *Edge FP*=extra edges added spuriously; *Edge FN*=edges present in gold but omitted.

| Model | # Evaluations | Average F1 (↑) |
|---|---|---|
| Llama-8B | 35 | $0.315 \pm 0.187$ |
| o3-mini | 35 | $0.452 \pm 0.208$ |

Table 11: Model Performance on Recent Papers (2024 Onwards)

| Model | # Evaluations | Average F1 (↑) |
|---|---|---|
| Llama-8B | 229 | $0.300 \pm 0.203$ |
| o3-mini | 235 | $0.410 \pm 0.199$ |

Table 12: Model Performance on Older Papers (Before 2024)

## L Degree of Confounding Labeling Prompt

Degree of Confounding is an important attribute of samples. It can act as a natural measure of difficulty, as nodes that are explicitly mentioned are easier to identify than entirely unobserved confounders. We use *R1* (et al., 2025) to label the degree of confounding for each node in each sample when provide the ground truth and source text using the prompt below. To allow for fine-grained analysis of degree of confounding, we allow for three different levels of confounding for each node. (i) Explicit (it or a synonym of the node's name appears in the text), (ii) Implicit (the node implicitly or indirectly appears in the text), or (iii) Absent (the node does not appear in the text whatsoever). We iteratively reject and retry any answers that do not meet formatting requirements until we receive valid answers for every benchmark sample to ensure that all samples have confounding levels. We detail the prompt used for this below. These node-level labels are used as the basis for calculation of degree of confounding.

**Label Unobserved Confounders Prompt**

~~~~
You will be given a causal graph in economics and a source text. Your task is to label each node in the graph to determine its degree of explicitness in the text. For each node, there are three possible levels:
1. The node (or the concept behind it) is explicitly mentioned in the text
- This can be verbatim, or though use of a synonym
- It is sufficient to be mentioned in the text; it is irrelevant if it is mentioned to be in the causal graph or not
2. The node is mentioned indirectly or implicitly in the text.
3. The node is unmentioned in the text, even if related concepts are discussed

Be conservative when determining the degree of explicitness for each node. Output only the JSON code block with your answer, without commentary, reasoning, explanation, or any other text. You must include the name of each node in the graph verbatim, even when the graph is very large, or many nodes are highly related or seem redundant.

# Expected Output Format

```json
{
    "scores": {
      "first_node_name": int_score_1_2_or_3,
      "second_node_name": int_score_1_2_or_3,
        ...
       "last_node_name": int_score_1_2_or_3
    }
}
```

It is MANDATORY to critically and thoroughly examine each and *every* node in the causal graph one at a time. Explicitly think about each node (and its corresponding relationships where appropriate) individually, even when it seems redundant or unnecessary. Even if it is tedious, you MUST do this and not take shortcuts.
~~~~

## M Variable Correction

We use the following prompt to correct the raw variable names extracted after annotation using *o3-mini*. To ensure validity, we use code-based approaches to automatically reject and retry any answers where all old names did not appear

**Variable Correction Prompt**

```
You are a world-class economist. You will be given a causal loop diagram (CLD) in JSON format. Your task is to combine variables that are intended to be the same, but are not named identically due to annotation errors. You will do this by combining variables and choosing which variable name to keep.

Your task is NOT to functionally alter the CLD. Be careful to only combine variables that are intended to be the same and are different solely due to annotation errors. When in doubt, do not combine the variables. Follow these guidelines:
- Avoid combining variables that are intended to be separate.
- Avoid combining variables that are highly similar but have different names.
- Do not create new variables or variable names, nor remove any variables from the CLD.
- Use the context of the CLD when making your decision.
- You must choose an existing variable name or your response will be rejected.
- You must only combine variables that are intended to be the same.
- Combining variables with more than one character difference between them is only done very rarely.

Positive examples:
- "Number of dog" and "Number of dogs" should be combined into "Number of dogs".
- "number of dogs" and "Number of dogs" should be combined into "Number of dogs".
Negative examples (do not combine):
- "<variable>" and "variable" should not be combined since it is clear that they are intended to be distinct.
- NEVER change any variables with < or > in the name.
- "Number of dogs" and "Number of hounds" should not be combined since it is clear that this isn't from an annotation error.
- "GDP" and "GNP" should not be combined; while they are only one letter apart, they are distinct variables.

Respond with your answer in JSON format and no other text.
JSON format:
{ "combined_variables": [
    {
        "old_names": ["variable1",
                      "variable 1",
                      "Variable1"],
        "new_name": "Variable1"
    },
    {
        "old_names": ["variable2",
                      "Variable 2",
                      "variable two"],
        "new_name": "Variable2"
    }
] }
```

## N PDF-to-Markdown Conversion

We utilize a LLM for the task of converting the text of the PDF to well-structured markdown as papers do not follow a consistent format. We find that reasoning models struggle at this task, and frequently fail to follow instructions to output the entire document by leaving out large sections of the text. We note that the normalization tool cannot be used for this task, as the numerous formatting errors and in-line citations would require it to be called once for almost every line of the text, and would result in an output many times longer than the source text. *Mistral Small* (Team, 2025a) follows the conversion instructions at tractable computational costs. We remove non-textual elements as they would be difficult to accurately represent in markdown. We additionally exclude irrelevant elements such as publication information and references as they are unrelated the economics task and needlessly inflate the length of texts. We also remove appendices, which are usually irrelevant or contain explicit information about the causal graph.

**PDF to Markdown Prompt**

```
Your task is to perform the minimal PDF pre-processing necessary to convert the provided PDF into a well-structured md file. Follow the guidelines below in order of priority:
1. Modify the text only when absolutely necessary. The exact wording of the original paper must be preserved verbatim.
- Do not correct spelling or grammar, even if it is incorrect
- The response will be rejected if even a single word is edited or removed unnecessarily; most of the response should effectively be copy-pasted from the original text
- Your response will likely be extremely long, around the same length as the original text; this is expected and normal.
2. Correct any broken text from the PDF processing and convert it into a well-structured md file.
- Convert sections and sub-sections into headings and subheadings
3. Remove the following information in entirety:
- Images, figures, and any other visual elements
- References and Citations, including when in-line. E.g., "[20, 22]" would be removed.
- Acknowledgments
- Authorship information
- Appendices
- Page numbers

Remember; your only output is the processed text in full, with no thinking, reasoning, or other commentary.
```

## O Text Normalization Prompt

In order to ensure the realism of the ReCAST benchmark, it is important to remove any explicit references to the causal graph, which make the task trivial. During this step, we also correct any references to non-existent elements which were removed in previous pre-processing steps (for example, referencing an image). We utilize a normalization tool to make these changes, which helps address several limitations of current LLMs. First, they struggle to output a large text in full, and have significant computational costs when doing so. Additionally, when outputting large chunks of text, they are prone to hallucinations and excessive edits, which are inappropriate. Using a tool also allows us to use code to check that the changes are valid; that is, that the start and end text are actually present in the text. We note that LLMs often struggle to account normalizations that overlap, even with specific prompting for this. In this case, a normalization will fail, and the entire response will be rejected. We iteratively prompt with the normalization prompt, stopping only when no normalizations are given. This ensures that it is confident that the text was correctly changed, and that no new text was introduced that needs to be changed. We utilize o3-mini (OpenAI, 2025) to perform this task, as it was shown to perform well during manual evaluation.

**Normalize Text Prompt**

Your task is to edit a md version of a published economics paper in markdown format to remove specific types of content.
- Remove any information that explicitly references the causal graph and its contents, including the causal graph itself
- This is the only information you should remove from the paper
- Only modify the text when it is necessary to remove the causal graph's information
- Only remove explicit references to the causal graph's elements, such as variable names, feedback loops, arrow colors, a variable explicitly being included, etc. Do not remove other references and related information to the causal graph, such as discussing elements of the causal graph, its relationships generally, and similar information
- You can only edit the paper; do not attempt to edit the causal graph
- The graph is supplied as a reference only in <causal_graph> tags
- Do not attempt to edit anything before </causal_graph>; this is not part of the paper and will be rejected

You have access to a special tool called 'normalize' that can replace text. This is the only way you can modify the text. Be careful to ensure that the text you are replacing is only the causal graph's information, and that it exists verbatim in the text.

The normalize tool takes three parameters:

1. start_string: The beginning of the text to replace
2. end_string: The end of the text to replace
3. replacement: The text to insert instead

You can call normalize multiple times to make several targeted replacements in the document. All three parameters are required for each call.

- By default, normalize will locate the *first* occurrence of the start_string. As a workaround for when the same text appears verbatim multiple times, use a slightly longer start_string and include some of the original text in your replacement to maintain context.
- Do not "redact" the text; remove references entirely rather than replacing them with generic text.
- Both the start and end strings will be included in the text that gets replaced. Changes are applied in order, so ensure that any string you replace is not used in another replacement or an error will be thrown.

Respond only with JSON in the following format:

```
{"normalizations": [
    {"start": "text to find (beginning)",
     "end": "text to find (end)",
    "replacement": "text to insert instead"},
     ...
  ]}
```

# P Causal Graph Generation Prompt

**Causal Graph Generation Prompt**

You are an expert causal reasoner and economist. Your task is to generate a causal graph for the provided markdown text. First, use extremely long chain-of-thought reasoning in <think> tags. Then, provide your final answer in a JSON code block, strictly following the following format:

```
```json
{
    "relationships": [
        {"source": causal_variable0,
         "sink": affected_variable0},
        {"source": causal_variable1,
         "sink": affected_variable1},
        ...
    ]
}
```
```

Your graph will contain exactly NUM_NODES nodes. When answering, do not provide any additional reasoning, commentary, or other information - only provide the JSON code block, with each dictionary representing one relationship in the graph.

# Q Standard Formatting Correction Prompt

**Formatting Correction Prompt**

Your task is to correct the formatting of a misformatted response, which is intended to end with a causal graph in economics that conforms to the proper JSON format. You will convert their intended answer to the proper JSON format, taking great care to be as faithful to the ground truth as possible. Do not attempt to modify the substance of their answer in any form, even if you think it may improve it's quality (including typos) - the task is to make the minimal changes possible to correct the formatting. The extent of the formatting may be minor, or be so extensive as to require writing the JSON from scratch.

Expected output format:

```
```json
{
    "relationships": [
        {"source": causal_variable0,
         "sink": affected_variable0},
        {"source": causal_variable1,
         "sink": affected_variable1},
        ...
    ]
}
```
```

You will be provided the original, misformatted answer. If it included lengthy intermediate steps, you will be given a snippet of them as context. Use only the final answer, always prioritizing the information provided closest to the end of the response.

If there is no text in the answer that resembles a causal graph, return an empty list of relationships.

Begin your response with the start of the JSON code block. Do not provide any reasoning, thinking, commentary, etc. - just the reformatted response. Don't overthink it.

## R Name-Assisted Causal Graph Generation

**Causal Graph Generation with Node Names Prompt**

You are an expert causal reasoner and economist. Your task is to generate a causal graph for the provided markdown text. First, use extremely long chain-of-thought reasoning in <think> tags. Then, provide your final answer in a JSON code block, strictly following the following format:

```json
{
    "relationships": [
        {"source": id_of_source_node,
         "sink": id_of_sink_node},
        {"source": id_of_source_node,
         "sink": id_of_sink_node},
        ...
    ]
}
```

You will be provided with the source markdown text and the name of each node in the graph. Ensure that each node is included at least once in the generated causal graph. Do not use the node's name in the graph; instead, use the id corresponding to the node. For the example nodes below (not the same as the ones you will be provided), whenever you want to include the node named "demand" in your graph, you would use the integer 2 rather than the word demand.

```json
{
    "nodes": [
        {"name": "supply", "id": 1},
        {"name": "demand", "id": 2},
        ...
    ]
}
```

When answering, do not provide any additional reasoning, commentary, or other information - only provide the JSON code block, with each dictionary representing one relationship in the graph.

Here are the nodes for your graph:

```json
NODE_JSON
```

## S Name-Assisted Formatting Correction Prompt

**Name-Assisted Formatting Correction Prompt**

Your task is to correct the formatting of a misformatted response, which is intended to end with a causal graph in economics that conforms to the proper JSON format. You will convert their intended answer to the proper JSON format, taking great care to be as faithful to the ground truth as possible. Do not attempt to modify the substance of their answer in any form, even if you think it may improve it's quality (including typos) - the task is to make the minimal changes possible to correct the formatting. The extent of the formatting may be minor, or be so extensive as to require writing the JSON from scratch.

In the original creation step, they were given the node names for the graph, each with corresponding ids. When correcting the graph, only ever use the integer ids corresponding to the node name, regardless of if the original used the names or correctly used the ids.

Expected output format:

```
{
    "relationships": [
        {"source": id_of_source_node,
         "sink": id_of_sink_node},
        {"source": id_of_source_node,
         "sink": id_of_sink_node},
        ...
    ]
}
```

You will be provided the original, misformatted answer. If it included lengthy intermediate steps, you will be given a snippet of them as context. Use only the final answer, always prioritizing the information provided closest to the end of the response. If it never comes to an answer, do not attempt to solve it yourself. Instead, simply return an empty list of relationships.

Begin your response with the start of the JSON code block. Do not provide any reasoning, thinking, commentary, etc. - just the reformatted response. Don't overthink it.

Here are the nodes for your graph:

```json
NODE_JSON
```

# T LLM-as-a-Judge Prompt

**LLM-as-a-Judge Prompt**

You are an expert economist. Your task is to act as an evaluator for a causal graph. You are provided with the ground-truth graph, the source text, and the LLM's response. You will also be told the type of evaluation to perform; only evaluate the response for that type of evaluation by closely following the instructions. Do not evaluate using any other type of evaluation.

When evaluating, follow these guidelines:
1. Follow each direction carefully, completely, and in-order
a. It is very important to be thorough and not take shortcuts, even when it seems tedious, redundant, or unnecessary. Do this for each node or edge you are evaluating; there is no time limit. Be sure to fully to fully think through each node or edge you are tasked with evaluating fully before moving onto the next one.
i. It is helpful to quote supporting evidence from the provided texts and graphs before reasoning about their relevance to the final evaluation for that node or edge.
ii. While evaluating a node or edge, you may examine several plausible counterparts to judge presence, semantics, abstraction, etc. (e.g., to see if it is broader or narrower than any ground-truth items). Use all relevant comparisons to inform your decision, but output one—and only one—set of labels for the item.
b. Only focus on the specific type of evaluation you are asked to do. Regardless of the accuracy (or lack thereof) in other categories, if you are asked to evaluate node precision, only evaluate node precision, not recall or edges. These are intended to be separate evaluations, so do not conflate the two.
c. Not Applicable labels must be explicitly selected when a category is skipped due to prior labels
d. Be conservative when grading - When in doubt between two labels, ere on the side of being harsh.

Start by thinking step-by-step in <think> tags. Then, output your answer in a YAML code block, formatted exactly as specified in the expected output format.

# Node Level Evaluation

## Node Precision
For each node in the LLM's response, evaluate against both ground truth sources:

1. Ground-Truth Graph Evaluation
- Explicitly identify and quote ALL potentially corresponding nodes from ground-truth graph
- Apply these labels where applicable:
Presence Labels (select one):
- PRESENCE_STRONG_MATCH: Core concept matches a ground-truth node with only minor, inconsequential differences
- PRESENCE_WEAK_MATCH: Core concept shares meaning with a ground-truth node, even if there are noticeable differences
- PRESENCE_NO_MATCH: There is no ground-truth node that captures a remotely similar core concept

Semantic Labels (select one):
- SEMANTIC_STRONG: Exactly or nearly identical meaning with only subtle distinctions
- SEMANTIC_MODERATE: Same core concept but with meaningful differences in scope or implication
- SEMANTIC_WEAK: Shares some semantic space but with substantial differences
- SEMANTIC_NA: Not applicable

Abstraction Labels (select one):
- ABSTRACTION_BROADER: Represents a more general concept that includes the ground-truth node
- ABSTRACTION_ALIGNED: Represents approximately the same scope and specificity of the ground-truth node
- ABSTRACTION_NARROWER: Represents a more specific subset of the ground-truth node
- ABSTRACTION_NA: Not applicable or the concepts were so different as to make abstraction comparison impossible

2. Ground-Truth Text Evaluation
- Explicitly quote ALL relevant supporting text from source
- Apply these labels where applicable:
Evidence Labels (select one):
- PRESENCE_STRONG_MATCH: Core concept appears in text with only minor, inconsequential differences
- PRESENCE_WEAK_MATCH: Core concept shares significant meaning with text but has notable differences
- PRESENCE_NO_MATCH: No text segments capture a similar core concept

Semantic Labels (select one):
- SEMANTIC_STRONG: Captures precisely what is stated in text or represents meaning with minimal interpretation
- SEMANTIC_MODERATE: Requires some interpretation but maintains core meaning
- SEMANTIC_WEAK: Significant interpretation needed; meaning partially preserved
- SEMANTIC_NA: Not applicable

Abstraction Labels (select one):
- ABSTRACTION_BROADER: Represents a more general concept that includes text concepts
- ABSTRACTION_ALIGNED: Represents approximately the same scope and specificity as the text
- ABSTRACTION_NARROWER: Represents a more specific subset of text concepts

- ABSTRACTION_NA: Not applicable or the concepts were so different as to make abstraction comparison impossible

## Node Level Recall
For each node in the ground-truth graph, evaluate against the LLM's response:

Response Evaluation
- Explicitly identify and quote ALL potentially corresponding nodes from LLM's response
- Apply these labels where applicable:
Importance Labels (select one):
- IMPORTANCE_CORE: Ground-truth node represents a fundamental concept central to the causal structure
- IMPORTANCE_INTERMEDIATE: Ground-truth node serves as a key connection between central concepts
- IMPORTANCE_PERIPHERAL: Ground-truth node provides supplementary or contextual information

Presence Labels (select one):
- PRESENCE_STRONG_MATCH: Core concept appears in response with only minor, inconsequential differences
- PRESENCE_WEAK_MATCH: Core concept shares significant meaning with a response node but has notable differences
- PRESENCE_NO_MATCH: No response node captures a similar core concept

Semantic Labels (select one):
- SEMANTIC_COMPLETE: Ground-truth concept fully captured with high fidelity, whether in single or multiple nodes
- SEMANTIC_PARTIAL: Core aspects captured but with some meaning loss or missing implications
- SEMANTIC_MINIMAL: Only basic or surface-level aspects of the concept captured
- SEMANTIC_NA: Not applicable

Abstraction Labels (select one):
- ABSTRACTION_BROADER: Represents a more general concept that includes the ground-truth node
- ABSTRACTION_ALIGNED: Represents approximately the same scope and specificity of the ground-truth node
- ABSTRACTION_NARROWER: Represents a more specific subset of the ground-truth node
- ABSTRACTION_NA: Not applicable or the concepts were so different as to make abstraction comparison impossible

# Edge Level Evaluation

## Edge Precision
For each edge (causal relationship) in the LLM's response, evaluate against both ground truth sources:

1. Ground-Truth Graph Evaluation
- Explicitly identify and quote ALL potentially corresponding edges from ground-truth graph
- Apply these labels where applicable:
Presence Labels (select one):
- PRESENCE_STRONG_MATCH: Edge connects highly similar concepts as in ground-truth
- PRESENCE_WEAK_MATCH: Edge connects somewhat similar concepts as in ground-truth
- PRESENCE_NO_MATCH: No corresponding edge exists in ground-truth

Directionality Labels:
- DIRECTION_CORRECT: Direction of causality matches ground-truth
- DIRECTION_REVERSED: Direction of causality is opposite of ground-truth
- DIRECTION_NA: Not applicable or the concepts were so different as to make direction comparison impossible

Abstraction Labels:
- ABSTRACTION_ALIGNED: Edge represents similar scope of relationship as ground-truth
- ABSTRACTION_BROADER: Edge is substantially more general than ground-truth
- ABSTRACTION_NARROWER: Edge is substantially more specific than ground-truth
- ABSTRACTION_NA: Not applicable or the concepts were so different as to make abstraction comparison impossible

2. Ground-Truth Text Evaluation
- Explicitly quote ALL relevant supporting text that describes causal relationships
- Apply these labels where applicable:
Evidence Labels (select one):
- PRESENCE_GRAPH_ONLY: Causal relationship present in ground-truth graph (always select this if present)
- PRESENCE_EXPLICIT: Causal relationship directly stated in text (only if not in graph)
- PRESENCE_IMPLIED: Causal relationship can be reasonably inferred from text (only if not in graph)
- PRESENCE_NO_MATCH: No text supports this causal relationship (only if not in graph)

Inference Labels (select one):
- INFERENCE_DIRECT: Relationship matches text's explicit causal claims
- INFERENCE_DERIVED: Relationship logically follows from text
- INFERENCE_STRETCHED: Relationship possible but weakly supported
- INFERENCE_NA: Not applicable or relationship does not exist

Abstraction Labels (select one):
- ABSTRACTION_ALIGNED: Matches the granularity of text's causal claims
- ABSTRACTION_BROADER: Generalizes multiple textual relationships
- ABSTRACTION_NARROWER: Specifies a subset

```
of text's causal claims
- ABSTRACTION_NA: Not applicable or the concepts were so different as to make abstraction comparison impossible

## Edge Level Recall
For each causal relationship (edge) in the ground-truth graph, evaluate against the LLM's response:

Response Evaluation
- Explicitly identify and quote ALL potentially corresponding causal relationships from LLM's response
- Apply these labels where applicable:
Importance Labels (select one):
Importance is based on how important it is to the ground-truth graph, regardless of whether it is present or accurately represented in the LLM's response.

- IMPORTANCE_CENTRAL: A key causal relationship that drives main effects
- IMPORTANCE_CONNECTING: Links major causal chains together
- IMPORTANCE_AUXILIARY: Provides supplementary causal context

Presence Labels (select one):
- PRESENCE_STRONG_MATCH: Core concept appears in response with only minor, inconsequential differences
- PRESENCE_WEAK_MATCH: Core concept shares significant meaning with a response node but has notable differences
- PRESENCE_NO_MATCH: No response node captures a similar core concept

Directionality Labels (select one):
- DIRECTION_CORRECT: Causal relationship captured with correct direction
- DIRECTION_REVERSED: Causal relationship present but direction is reversed
- DIRECTION_UNCLEAR: Relationship present but direction is ambiguous
- DIRECTION_MISSING: Relationship entirely absent from response

Abstraction Labels (select one):
- ABSTRACTION_ALIGNED: One-to-one relationship match at similar level of detail
- ABSTRACTION_BROADER: Edge is substantially more general than ground-truth
- ABSTRACTION_NARROWER: Edge is substantially more specific than ground-truth
- ABSTRACTION_NA: Not applicable or the concepts were so different as to make abstraction comparison impossible

# Expected Output Format
The output should be in YAML format. Only include the evaluation sections that are being evaluated - omit other sections entirely. For example, if only evaluating node precision, only the node_precision_evaluations section should be present. However, within the required evaluation sections, be sure to always include the Not Applicable labels rather than omitting them.

"'yaml
# If evaluating node precision:
node_precision_evaluations:
- node_number: <integer>
graph_evaluation:
presence_label: <PRESENCE_LABEL>
semantic_label: <SEMANTIC_LABEL>
abstraction_label: <ABSTRACTION_LABEL>
text_evaluation:
presence_label: <PRESENCE_LABEL>
semantic_label: <SEMANTIC_LABEL>
abstraction_label: <ABSTRACTION_LABEL>

# If evaluating node recall:
node_recall_evaluations:
- node_number: <integer>
importance_label: <IMPORTANCE_LABEL>
presence_label: <PRESENCE_LABEL>
semantic_label: <SEMANTIC_LABEL>
abstraction_label: <ABSTRACTION_LABEL>

# If evaluating edge precision:
edge_precision_evaluations:
- edge_number: <integer>
graph_evaluation:
presence_label: <PRESENCE_LABEL>
directionality_label: <DIRECTION_LABEL>
abstraction_label: <ABSTRACTION_LABEL>
text_evaluation:
presence_label: <PRESENCE_LABEL>
inference_label: <INFERENCE_LABEL>
abstraction_label: <ABSTRACTION_LABEL>

# If evaluating edge recall:
edge_recall_evaluations:
- edge_number: <integer>
importance_label: <IMPORTANCE_LABEL>
presence_label: <PRESENCE_LABEL>
directionality_label: <DIRECTION_LABEL>

abstraction_label: <ABSTRACTION_LABEL>
"'
```

## U LLM-as-a-Judge Scoring Mechanics

The quantitative metrics derived from the LLM-as-a-Judge's YAML output are calculated as follows. First, the judge's qualitative labels for various evaluation criteria (e.g., PRESENCE_STRONG_MATCH, SEMANTIC_MODERATE, IMPORTANCE_CORE) are mapped to pre-defined numerical scores, ranging from 0.0 (no match/irrelevant) to 1.0 (perfect match/highly important). For multi-faceted evaluations like node precision, which considers presence, semantic similarity, and abstraction level, a composite score for a single aspect (e.g., node precision against the ground-truth graph) is computed by averaging the numerical scores of its constituent labels.

Precision metrics (node precision, edge precision) for each item generated by the LLM are determined by comparing it against both the ground-truth graph and the source text. If the item is labeled as `PRESENCE_NO_MATCH` against both sources, its score is 0.0. Otherwise, the higher of the two composite scores (one from graph comparison, one from text comparison) is taken as the item's precision score. The overall precision for a category (e.g., node precision) is then the arithmetic mean of these individual item precision scores.

Recall metrics (node recall, edge recall) assess how well the LLM's output captures items from the ground-truth graph. For each ground-truth item, a composite correctness score is calculated based on its presence and the fidelity of its representation in the LLM's output (considering factors like semantics, abstraction, and directionality for edges). This correctness score is then multiplied by a numerical `importance` weight assigned by the judge to that ground-truth item (e.g., `IMPORTANCE_CORE` receives a higher weight than `IMPORTANCE_AUXILIARY`). The final recall score for a category is a weighted average: the sum of (correctness score $\times$ importance weight) for all ground-truth items, divided by the sum of all possible importance weights. This ensures that correctly recalling more important ground-truth items contributes more significantly to the recall score.

Finally, F1 scores for nodes, edges, and overall performance are calculated using the standard harmonic mean: $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$. Overall precision and recall are micro-averaged, where the total weighted sum of correct predictions is divided by the total number of predictions (for precision) or total ground-truth items (for recall, considering importance weights), across both nodes and edges.